# Towards Ubiquitous Personalized Music Recommendation with Smart Bracelets

JIAYU LI, DCST, Tsinghua University, China
ZHIYU HE, DCST, Tsinghua University, China
YUMENG CUI, DCST, Tsinghua University, China
CHENYANG WANG, DCST, Tsinghua University, China
CHONG CHEN, DCST, Tsinghua University, China
CHUN YU, DCST, Tsinghua University, China
MIN ZHANG*, DCST, Tsinghua University, China
YIQUN LIU, DCST, Tsinghua University, China
SHAOPING MA, DCST, Tsinghua University, China

Nowadays, recommender systems play an increasingly important role in the music scenario. Generally, music preferences are related to internal and external conditions. For example, mood state and ongoing activity will affect users' music preferences. However, conventional music recommenders cannot capture these conditions since they only utilize the online data but ignore the impact of physical-world information. In this paper, we leverage the contexts from low-cost smart bracelets for ubiquitous personalized recommendation to meet users' music preference. We first conduct a large-scale questionnaire survey, which illustrates moods, activities, and environments will affect music preferences. Then we perform a one-week field study among 30 participants, where they receive personalized music recommendation and record preferences and mood. Meanwhile, participants' context information is collected with bracelets. Analyses on the data demonstrate significant relationships between music preference, mood, and bracelet contexts. Furthermore, we propose a novel Multi-task Ubiquitous Music Recommendation model (MUMR) to predict personalized music preference with bracelet contexts as input and mood prediction as an auxiliary task. Experiments show significant improvement in music recommendation performances with MUMR. Our work demonstrates the possibility of ubiquitous personalized music recommendations with smart bracelets data, which is an encouraging step towards building recommender systems aware of physical-world contexts.

CCS Concepts: • **Information systems** → **Recommender systems**; • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Personalized ubiquitous recommendation, context-aware music recommendation, field study.

*Min Zhang is the corresponding author. z-m@tsinghua.edu.cn

Authors' addresses: Jiayu Li, DCST, Tsinghua University, 30 Shuangqing Rd, Haidian Qu, Beijing Shi, China, jy-li20@mails.tsinghua.edu.cn; Zhiyu He, DCST, Tsinghua University, 30 Shuangqing Rd, Haidian Qu, Beijing Shi, China; Yumeng Cui, DCST, Tsinghua University, 30 Shuangqing Rd, Haidian Qu, Beijing Shi, China; Chenyang Wang, DCST, Tsinghua University, 30 Shuangqing Rd, Haidian Qu, Beijing Shi, China; Chong Chen, DCST, Tsinghua University, 30 Shuangqing Rd, Haidian Qu, Beijing Shi, China; Chun Yu, DCST, Tsinghua University, 30 Shuangqing Rd, Haidian Qu, Beijing Shi, China; Min Zhang, z-m@tsinghua.edu.cn, DCST, Tsinghua University, 30 Shuangqing Rd, Haidian Qu, Beijing Shi, China; Yiqun Liu, DCST, Tsinghua University, 30 Shuangqing Rd, Haidian Qu, Beijing Shi, China; Shaoping Ma, DCST, Tsinghua University, 30 Shuangqing Rd, Haidian Qu, Beijing Shi, China.

2474-9567/2022/9-ART125
https://doi.org/10.1145/3550333

**ACM Reference Format:**
Jiayu Li, Zhiyu He, Yumeng Cui, Chenyang Wang, Chong Chen, Chun Yu, Min Zhang, Yiqun Liu, and Shaoping Ma. 2022. Towards Ubiquitous Personalized Music Recommendation with Smart Bracelets. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 125 (September 2022), 34 pages. https://doi.org/10.1145/3550333

## 1 INTRODUCTION

With the rapid growth in the digital music industry, recommender systems play an increasingly important role in the music scenario [59, 66]. Generally, people prefer different music under different contexts, such as environment, mood, and time [4, 23, 61]. For instance, depressed people tend to prefer sad music with low energy [75], and distinct playlists are favored in different seasons [38]. However, most context-aware music recommender systems tried to obtain the user-generated contexts from online interactions [48, 48, 58, 60], which is not a real-time reflection of user status in the physical world. With the development of the Internet of Things (IoTs) and ubiquitous computing, wearable devices provide a new way to collect physical-world context. They have been widely used for users' physical and mental status prediction [25, 33, 41, 73], but little attention was focused on the personalized recommendation with wearable devices.

Nevertheless, performing personalized ubiquitous music recommendation with wearable devices is challenging. As suggested by Cross [17], music preference judgment is a high-level mental process involving complex cognitive behaviors. However, wearable devices, such as smartwatches and smart bracelets, provide only elementary contexts about the environment and low-level physical signals. Therefore, a large cognitive gap exists between music preferences and context data from wearable devices. Furthermore, wearable devices collect limited context, and it is questionable whether they are enough to reflect people's complex daily activities influencing music preference.

Considering the above challenges, in this work, we try to answer the following question: *Is it possible to perform ubiquitous personalized music recommendation with contexts collected by low-cost wearable devices?* Low-cost commercial smart bracelets are adopted because they collect primary contexts and are widely used in daily life. With the low-cost bracelets, three main categories of context can be collected: biological signals (e.g., heart rate), activity (e.g., activity type and intensity), and environment (e.g., location, time, and weather). We aim to analyze the influence of these real-world contexts on user preferences in interactions between users and music recommender systems, and further perform ubiquitous personalized recommendation with real-world contexts.

To achieve the above goals, we analyze the influence of real-world contexts on music preferences with questionnaires, a field study, and a novel recommender. Firstly, we conduct a large-scale questionnaire with 350 participants to analyze contexts' influence on music preference in daily life. The answers show that music preference is widely influenced by listeners' moods, activities, and environment. Especially, the subjective status, mood, has a strong influence on music preference. Since existing works have revealed that mood can be detected with context [32, 33, 40, 41], we treat it as an important latent influence factor. Inspired by findings from the questionnaires, we conduct a user study with 30 participants for ubiquitous music recommendation. It includes a one-week field study to collect participants' contexts with smart bracelets in daily life. Meanwhile, participants are required to receive music recommendations from the server, record preference ratings, and mark their mood before and after music listening. Analyses of the participants' context and music feedback further illustrates that music preferences are associated with moods, activities, environments, and biological signals.

In the end, we propose a Multi-task Ubiquitous Music Recommendation (MUMR) model to perform personalized recommendation with bracelet-collected contexts and user-labeled mood. User, music, and bracelet contexts are fused as input for preference prediction, and mood serves as an auxiliary prediction goal in a multi-task optimization strategy. Experiments on the field study dataset show significant performance improvement with MUMR. We also perform a 1-week real-world testing of MUMR with 10 participants from the former user study. Compared with methods without bracelet contexts, MUMR shows significant improvement, which further indicates the usefulness of bracelet contexts in music recommendation.

Therefore, it is possible to perform ubiquitous personalized music recommendation with context information collected by low-cost bracelets, even without explicit annotations from users. The main contributions of this work are as follows:

- We reveal a direction of ubiquitous personalized music recommendation with low-cost wearable devices. A field study is conducted to understand and perform personalized music recommendation in an uncontrolled environment with bracelet collected context.
- We discuss the influence of rich context on music preference, including mood, environment, activity, and biological signals. Both large-scale questionnaires and user studies show significant relationships between contexts and music preferences. Especially, user mood has a strong impact on music preference as an important latent factor.
- We propose a Multi-task Ubiquitous Music Recommendation model (MUMR) for ubiquitous personalized music recommendation, with bracelet context as input and mood prediction as an auxiliary task. Both offline experiments and real-world testing of MUMR show promising performances on music preference prediction and recommendation.
- After removing sensitive information, we make the field study dataset open source at https://github.com/JiayuLi-997/MUMR-Ubiquitous_Recommendation. To the best of our knowledge, this is the first public dataset for ubiquitous recommendation in the wild.

The remainder of this paper is organized as follows: We review the related work in Section 2, and present findings from a large-scale questionnaire in Section 3. The user study methodology is shown in Section 4. Section 5 provides analyses on data collection of the user study. In Section 6, we present the personalized recommender model design, experimental settings, and results. And we discuss our main concerns and limitations in Section 7. Finally, Section 8 provides the conclusions of our work.

## 2 RELATED WORK

In this section, we introduce the music recommendation scenario, general context-aware recommendation methods, location-based ubiquitous recommendation, and the progress of wearable devices for detection tasks.

### 2.1 Music Recommender Systems

Music is one of the most important scenarios for recommender systems. As users' music preferences are dynamic and diverse, music recommendation relies on context attributes in nature [52, 59].

In conventional music recommendation, music audio signals and descriptive metadata were used as item-level content [27, 56, 66, 76]. As more user profiles can be collected online, user-generated contexts have been considered to improve the recommendation performance. In [14], the characteristics of venues and music were mapped into a latent semantic space to measure the similarity between locations and music. Emotion-oriented user features were extracted from the user blog for music recommendation in [60]. The social context was also considered, such as online social networks [48], social influence [11], and cultures of users [58, 77]. However, these works focused on the online context data, which is often not real-time. Instead, we perform ubiquitous recommendation with physical-world information to capture real-time changes in user status and music preferences.

In recent years, some attempts have been made to consider the real-world contexts and emotions in the music recommenders [4]. Some researchers used sensor signals directly for music recommendation. Elliott and Tomlinson [20] calculated the walking pace from accelerometer signals and recommended music with a tempo equal to the pace. Heart rates collected with bio-sensors were also used for selecting proper music to guide users to a target heart rate [15, 44, 53]. These researches have revealed a strong link between music and biological signals.

However, they all focused on aligning the music and psychological signals, which is not necessarily relevant to recommending satisfying music. Optimizing user satisfaction is more complex than adjusting physiological status because it involves high-order cognitive processes. Moreover, user satisfaction is a fundamental goal for recommenders [59] and also the aim of our research.

Some researchers tried to extract semantic information (e.g., music emotion or rhythm, and user activity or mood) from both the music and real-world signals, and then recommended by matching music to similar user statuses. Wang et al. [71] manually classified each music track to proper activities. Then they predicted the user's ongoing activity with smartphone sensors and recommended music in the activity class. Kim et al. [35] predicted music tempo levels with audio features and detected users' activities with accelerometer signals. Then general / user-specific rules were used to match music tempo levels with user activity intensity. By detecting the emotion of music and users, a rule-based method was utilized to recommend music with similar emotions to users [51]. In *DJ-Running* [2], music emotion was detected by audio features, and used to provoke similar emotions in runners to help them get through long-distance running. However, rule-based matching between music and user is not always effective. As revealed by our questionnaires, music with emotions different from users' status may also be welcome. Moreover, in previous works, either personalized recommendation was not considered, or the music was post-filtered by users' preferred music genres, which is insufficient to reflect personalized preference changing with context. Therefore, we abandon the idea of matching similar music and status, and instead conduct a novel personalized recommender to automatically learn the most appropriate music for users in different contexts.

In addition, the previous music recommenders with real-world contexts paid much attention to classifying user status (e.g., activity and mood) from the original signals (e.g., accelerometer, ECG, etc.). However, we focus on the recommendation strategy to satisfy users with proper music based on user status detected by low-cost bracelets. Therefore, we do not compare with existing real-world context recommenders in the experiments.

## 2.2 Context-aware Recommendation

Context-aware recommendation utilizes various information in the recommendation process to improve system performances, which results in a multidimensional task of modeling user preference in the space of user × item × context [1]. In conventional music recommenders, side information about items and user profiles has been widely used as input for deep neural networks, such as Wide& Deep [13], DeepFM [26], and DIN [80]. In recent works, the graph structure has been incorporated for modeling knowledge graph of items [62, 70], social networks among users [22], and multiple user behaviors [78]. Furthermore, temporal information has been incorporated for sequential recommendation [46, 68].

Contexts in the physical world were also considered in different recommendation scenarios. For instance, the spatial location was widely used for service recommendation to provide fast service for users [74, 79]. Location was also used to measure the similarity between users, and geographically close users were recommended the same items [69]. *Motivate* [43] was a context-aware mobile recommender system that detected user activity and promoted a healthy lifestyle, which provided 34 pieces of advice for physical activities with rule constrains. In [19], location, weather, and movement were collected by smartphone, and recommendation for activity was conducted with pre-defined rules.

However, these works utilized the same rules for all users, e.g., recommended a user to take a walk if it was a sunny weekday and her agenda was free after lunch [43]. Since the rules are manually pre-defined for the focused scenario (e.g., activity recommendation), they are not extensible to other applications, such as music recommendation. Therefore, we do not compare with these methods in our experiments. By conducting personalized ubiquitous music recommendation, we analyze whether and how various contexts influence preference and provide personalized recommendations for different users.

### 2.3 Ubiquitous Recommendation

In previous works, ubiquitous recommendation indicated that the recommender systems provided users with personalized recommendations of items in the proximity [47]. The applied scenarios were usually related to mobile tourism, such as shopping recommenders, tourist guides, and route finders [49]. For instance, during offline store shopping, NFC tags were used to identify nearby products for customers, and the detected product information was sent via smartphone in [55]. Similar methods for displaying product features or reviews were used in [30, 67]. CityVoyager estimated users' preferences from their historic locations and recommended the nearby shop in surroundings [63]. A mobile recommender system for tourist guides was proposed in [24], where points of interest (POIs), current location, and time are considered for recommending the next place to visit. In [12], surrounding hotel recommendation was proposed by location and explicit requirement from users.

However, all these ubiquitous recommendation systems focused on the location context and recommended nearby items to users, which limited the recommendation scope. With smart bracelets, we aim to collect and incorporate more contexts about user status and recommend online music resources regardless of physical user-item distance.

### 2.4 Wearable Devices in Human Status Detection

With the development of portable sensors, wearable devices have been widely applied for human status detection and prediction. One of the most fundamental tasks for wearable devices is to perform daily self-monitoring, such as diet control [7], smoking cessation [34], and treatment for chronic diseases [5]. In these works, records and preliminary analysis of behaviors helped users take healthy behaviors and lifestyles. Further, wearable sensors were used for physical status detection, such as behavior recognition and sleep quality detection [10, 25, 65], which have already been used in commercial devices. Detection and prediction of mental status was also a trend for wearable device applications. Multiple biological metric sensors were widely used for mood, stress, and well-being [32, 33, 50, 54, 81]. Li et al. [41] conducted a series of works on personality prediction, mood detection, and depression detection with low-cost wearable devices. More complex human status in certain conditions was also detected, such as work efficiency for employees [57] or study quality of students [18].

Various wearable devices were used in the previous works to detect physical and mental status of human beings. The encouraging results of these researches show the possibility to fill the gap between basic context signals and high-level cognitive tasks, which inspire us to model music preference with the bracelet data.

## 3 LARGE-SCALE QUESTIONNAIRES FOR CONTEXT INFLUENCE ON MUSIC PREFERENCES

Firstly, to motivate our experiment, we conduct a pre-study questionnaire to gain insight into influence factors for music preference in daily life.

### 3.1 Questionnaire Design

We design the questionnaire with a commercial online survey platform[1], and the content is shown in Appendix A. As we perform the experiments in China, all materials were originally presented in Chinese, and then translated to English for publication. After collecting demographic (Q1-Q4), the survey recalls participants' memory about recent music listening experiences with questions of music listening frequency and preferred music genres (Q5-Q7). Then participants are required to rank factors influencing music preference from a list of music features and real-world contexts (Q8). Finally, participants imagine themselves in different situations (i.e., moods and activities) and choose frequency, preferred genres, and purposes of music listening in each situation (Q9-Q20). The options for music genres and purposes are inspired by the questionnaires on activity and music by Song

[1] https://www.wjx.cn/

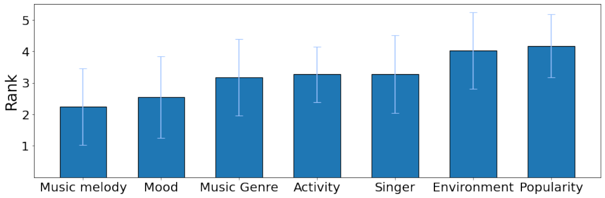


Fig. 1. Mean and standard deviation of participants' rank positions of the influence factors for music preference. A smaller rank indicates greater importance, and the error bar stands for the standard deviation.

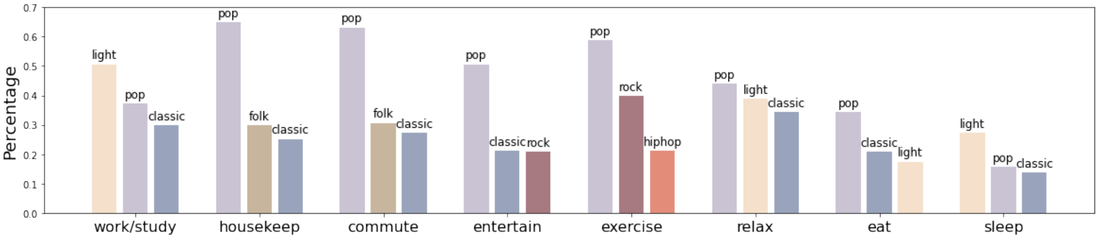


Fig. 2. Top-3 most favored music genres when participants engage in different types of activities.

[61]. In all ranking, matrix, and multiple-choice problems, the options (rows for matrix problems) are randomly shuffled for each participant to reduce bias from option positions or participant fatigue [8].

### 3.2 Participants

Participants were recruited with online BBS and social networks. In total, 389 responses were collected, and 350 of them were valid according to the attention test question (Q13). Most of the participants were students at a public university. The gender proportion is *female:male:others=213:135:2*. The age distribution is *Under 18: 18-25: Over 25 = 6:235:109*. All participants had listened to music on mobile Apps in the past month. And distribution of average music listening time per day is *less than 1 hour: 1-2 hours : more than 2 hours = 133:138:79*.

### 3.3 Analysis and Findings

Based on 350 valid responses, we inspect how various contexts influence music preference.

Firstly, we analyze the importance of various factors influencing music preference. We require participants to *sort the following factors according to how much they influence your music preference* (Q8). Among all factors, three are context-related (mood, activity, and environment), others are music-related. The average rank positions and standard deviations are shown in Figure 1, where a smaller rank indicates greater influence, and the error bar shows the standard deviation. Among all factors, *mood* ranks the second place, *activity* ranks the forth, and *environment* ranks the sixth. It demonstrates that contexts play an essential role in music preference, even more important than some attributes of music itself. Especially, *mood* has a significantly stronger influence than *genre*, *singer*, and *popularity* of music with paired t-test p-value<0.01, respectively. It inspires us to collect mood and discuss its interplay with music in the field study.

Furthermore, the music genre preferences in different contexts are analyzed. Firstly, we count the percentage of each genres selected in Q7, and find that the four most popular music genres are *pop* (83.1%), *folk/country* (48.0%), *classical* (47.1%), and *light* (45.4%). Then, participants choose their favored genres in different moods (Q11) and activities (Q20). Notice that among three context-related factors in Figure 1, we discuss the genre preference in

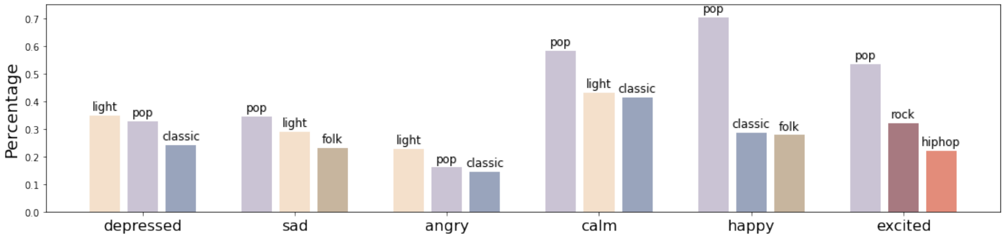


Fig. 3. Top-3 most favored music genres when participants are in different mood.

various *moods* and *activities*, whereas *environment* is left out, since it is difficult to imagine staying in different *environment*. Percentage of selecting each music genre is counted, and top-three most favored genres in each activity and mood are shown in Figure 2 and Figure 3, respectively. *Pop* is the only music genre showing in all activities and moods, while other genres are preferred in different situations. Considering music in different activities, *light* music is most favored while working/studying and sleeping, and *rock* is preferred for entertainment and sports. Comparing music preferences in diverse mood status, the music genres participants prefer when excited differ from those in other moods greatly. Passionate music is favored when excited, such as *rock* and *hip-hop*. Thus, music genre preferences are generally influenced by listeners' activity and mood.

We also ask participants about their *preferred music emotion in different moods* (Q12). We find that when feeling sad, the proportion of people willing to listen to quiet music (31.7%) is close to those who prefer sad music (30.0%). And *calm* music is welcome in excited status (31.4%), so is the opposite (25.7%). It indicates user mood and preferred music emotion are not identical, so simple recommendation method to match mood with similar music emotion may not always be proper.

Moreover, for the question *Do you think your moods will change before and after listening to music?* (Q15), 90.3% participants selected *Always*, *Often*, or *Sometimes*. So it is very common to experience mood changes during music listening.

The questionnaire survey reveals tight connections between music preference and contexts including mood, activity, and environment. Therefore, real-world contexts should be considered when performing music recommendation. Especially, mood is an important influence factor. Whereas, as a subjective element, it cannot be automatically collected. So mood is considered as a latent factor in the following experiments.

## 4 METHODOLOGY FOR USER STUDY ON UBIQUITOUS MUSIC RECOMMENDATION

Next, we conducted a user study to collect users' music preference and real-world contexts in daily life. At first, participants come to the laboratory, listen to some music tracks, and give preference feedback on each track. The feedback is used to train a personalized recommender for each participant. Then they wear a smart bracelet in a one-week field study, listen to recommended music everyday, and give preference and mood feedback. Participants can exit at any point and will be paid according to the number of completed music listening tasks. The research protocol was reviewed and approved by the Department of Psychology Ethics Committee, Tsinghua University (THU202118). In this section, we introduce methodology and procedures for our user study.

### 4.1 Participants

We recruited 32 participants from a public university and the surrounding community, and 2 participants dropped out of the experiment due to personal reasons. Among the remaining 30 participants, 18 were female, and 12

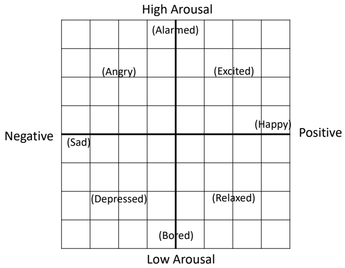


Fig. 4. Two-dimensional description map of Thayer mood model, where the X-axis represents valence, and the Y-axis represents arousal [64].

were male, aged 18-55 (M=24.13, SD=7.39[2]). 14 participants were undergraduates, 13 were graduates, 3 were faculty members of the university, and one was a company engineer. All participants were well informed of the experiment and signed the informed consents. To incentivize participants, we paid them by the number of tasks they finished [29]. The basic payment is $2.34 per day, and participants can gain up to $9.38 if they listen to seven or more music tracks per day.

## 4.2 Data Collection Preparation

*4.2.1 Music Selection.* We select 1000 music tracks from the MillionSong Dataset [6]. MillionSong contains a collection of audio features for a million contemporary popular music tracks until 2011[3], and a *Taste Profile Dataset*[4] of users' play counts on each track from 1 million users of Spotify. Firstly, we select a subset of music tracks. We divide the *music valence* feature into 10 equal intervals of all tracks and select the top-1000 most popular music tracks (in terms of total play counts in the user dataset) from each valence interval. We consider the valence feature to ensure our dataset contains various music emotions. Then, music tracks with *demo* or *live* in the title are excluded since their quality cannot be guaranteed. Finally, 9704 music tracks are left, and we randomly select 1000 tracks from them. Secondly, because an external dataset of interactions between users and music tracks is necessary for recommender training, we choose a subset from the *Taste Profile Dataset* [4], where users who have listened to at least 15 of the 1000 music tracks are selected. An interaction between user $u_i$ and music track $m_j$ is recorded if $u_i$'s play count on $m_j$ is greater than zero. Eventually, we obtain an external interaction dataset with 1000 tracks, 5957 users, and 115,365 interaction pairs $(u_i, m_j)$.

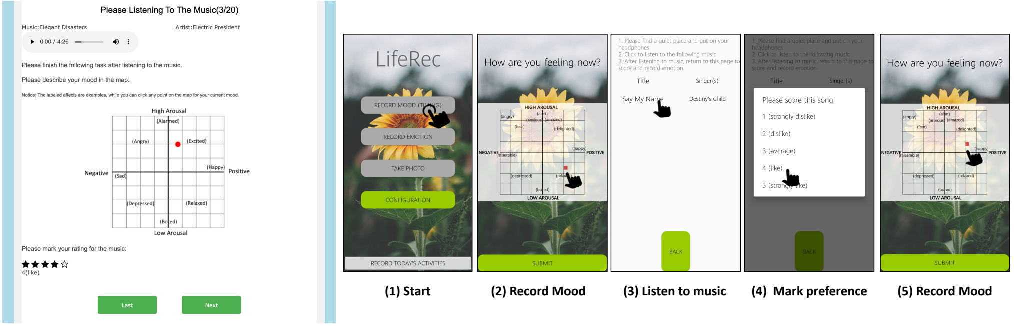
*4.2.2 Mood Labeling.* As we require participants to record mood frequently, an intuitive method for mood labeling is necessary. Balancing the label precision and complexity, we implement the Thayer mood model[64] with a two-dimensional description map of mood, i.e., valence (positive or negative) and arousal (energetic degree), as shown in Figure 4. During user study, participants only need to click a point on the map to represent their moods.

*4.2.3 Website and App Implementation.* In the lab study, participants were required to record their music preferences on the server, so a concise website interface was designed for music listening, preference rating, and mood labeling, as shown in Figure 5(a). In the field study, participants were instructed to use a public lifelog recording App, LifeRec [42], to record mood and receive music recommendations. Since Section 3.3 illustrates

[2] M for mean value, SD for standard deviation

[3] http://millionsongdataset.com/pages/getting-dataset/

[4] http://millionsongdataset.com/tasteprofile/



Please Listening To The Music(3/20)

Music:Elegant Disasters Artist:Electric President

Please finish the following task after listening to the music.

Please describe your mood in the map:

Notice: The labeled affects are examples, while you can click any point on the map for your current mood.

Please mark your rating for the music:

★★★★☆ 4(like)

1. Please find a quiet place and put on your headphones
2. Click to listen to the following music
3. After listening to music, return to this page to score and record emotion.

| Title | Singer(s) |
|---|---|
| Say My Name | Destiny's Child |

Please score this song:

1 (strongly dislike)

2 (dislike)

3 (average)

4 (like)

5 (strongly like)

(a) Lab Study Interface

(b) Field Study Interface

Fig. 5. Interface of the lab study music preference collection task and the field study music listening task on LifeRec App [42].

that it is prevalent to experience mood changes during music listening, we asked participants to record mood before and after listening to music, respectively. An example of a music listening task by LifeRec is shown in Figure 5(b), including five steps: (1)start, (2)record mood, (3)receive and listen to a recommended track, (4) record 5-level rating (i.e., preference), and (5) record mood again.

*4.2.4 Smart Bracelet and Data Transmission.* We select MiBand3[5] smart bracelet to record users' **environment**, **activity**, and **bio-metrics**. This bracelet is chosen because it is one of the most popular wearable devices in China, incredibly portable, and low-cost (about $15 for each). In the field study, we categorize the collected bracelet data into three types, **Envorinment**, **Activity**, and **Biological signals**. **Environment**, including GPS and corresponding weather information, was acquired every 15 minutes. Time of recommendation was also recorded as part of **Environment**. Steps, activity intensity, and activity kind were collected at 1Hz as **Activity**. The **biological signals**, i.e., heart rate, was also collected at 1Hz frequency. The bracelet data was transferred to the participants' smartphones with a public App GadgetBridge[6] automatically. For privacy concerns, data would not be uploaded to the server until it was checked and consented to upload by the participants at the end of experiment.

## 4.3 User Study Procedure

The overall illustration of user study procedure is shown in Figure 6.

*4.3.1 Lab Study Session (60 min).* Before the user study, 20 music tracks are randomly selected from the 1000-songs dataset for all participants. To start the experiment, Every participant is required to come to the lab and record music preference on the 20 tracks, which takes about 60 minutes (Left in Figure 6). Firstly, after signing the informed consent, the participant takes a pre-experiment interview about demographics and general music preferences. Then, we help the participant configure two Apps (i.e., GadgetBridge and LifeRec) and connect a smart bracelet to her smartphone. Configurations about GadgetBridge include allowing the App to run in the background, using the heart rate sensor to improve sleep detection, and opening whole-day heart rate measurement once a minute. Settings on Liferec include User ID and time to receive task notification. Afterwards, 20 tracks are shuffled and presented to each participant. And he/she needs to label preference rating and mood for each track on the website in Figure 5(a). Finally, participants will try a field study task on their smartphones following steps in Figure 5(b).

[5] https://www.mi.com/shouhuan3/

[6] https://codeberg.org/Freeyourgadget/Gadgetbridge/wiki

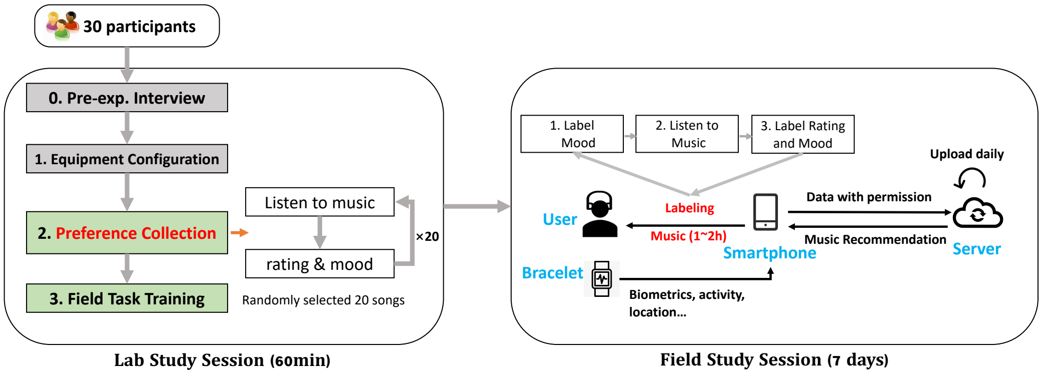


Fig. 6. Illustration of user study procedure. The user study includes two sessions: the Lab Study Session to collect participants' music preference, and the Field Study Session to collect smart bracelet data and feedback on daily music recommendation.

*4.3.2 Field Study Session (1 week).* Following the lab study, participants will start a one-week field study to collect data in daily life, as shown in right part in Figure 6. They are required to wear smart bracelets the whole day to automatically collect the information mentioned in Section 4.2.4. Meanwhile, every 1-2 hours, participants would be reminded to perform one music listening task on LifeRec. For music recommendation, in each task, a remote server would receive the participant's ID, and recommend a track with three recommenders in turn: *Random*, *Wide & Deep* [13], and *LightGCN* [28]. Three methods are selected in order to alleviate possible bias caused by a single recommender [45]. In *Random*, a random track will be recommended. As for *Wide & Deep* and *LightGCN*, two personalized models would be updated daily for each participant, based on the external dataset in Section 4.2.1 and personal preference ratings in both lab and field study. In the recommendation process, each participant will listen to a track only once.

During the field study, participants will be interviewed on the second day to ensure their data is recorded correctly. At the end of field study, the participants will come to the lab, inspect all records in detail, and consent to upload them to the server. Finally, they fill in the questionnaire in Appendix A, give preference to each of the genres in Q7 (in terms of *like*,*neutral*,*dislike*), and get paid by the number of records they have contributed during the experiment.

## 4.4 Data Pre-processing

Before analyzing and detecting, we pre-process the user, music, and bracelet context data.

Each **user** is represented with an anonymous participant ID, and each **music** track is represented with 20 audio features from Spotify[7] and ComParE2013 acoustic feature set[72], which are shown in Table 1.

As for bracelet data, we perform the following process for each category of records:

- **Environment**: Environment contains time, location, and weather. The record *time* is divided into 3 phases with equal-number records: 8a.m.-12p.m., 12p.m.-6p.m., and 6p.m.-2a.m. (No records are collected during 2a.m. to 8a.m.). The GPS records, i.e., longitude and latitude, closest (in time) to each music listening records are used as features for *locations*. For privacy concerns, location records are transferred into relative distance (meters) to the center of all locations in the dataset. We collect the geographical locations rather than semantic locations because the participants mostly have a regular schedule in or around the campus, and the algorithm is expected to automatically learn semantic information of the locations. The

[7] https://developer.spotify.com/console/get-audio-features-track/

Table 1. Twenty music features extracted from Spotify and ComParE2013 acoustic feature set, where *Org. Name* is the original feature name in Spotify/ComParE, and *Feature Name* is the name used in this paper.

| Source | Org. Name | Feature Name | Source | Org. Name | Feature Name |
|---|---|---|---|---|---|
| Spotify | popularity | popularity | Spotify | tags | genre[1] |
| Spotify | loudness | loudness | Spotify | duration | duration |
| Spotify | danceability | danceability | ComParE | F0final_sma_amean | pitch |
| Spotify | energy | energy | ComParE | F0final_sma_stddev | picth_std |
| Spotify | key | key | ComParE | lengthL1norm_sma_stddev | loudness_std |
| Spotify | speechiness | speechiness | ComParE | RMSenergy_sma_stddev | energy_std |
| Spotify | acousticness | acousticness | ComParE | psySharpness_sma_amean | sharpness |
| Spotify | instrumentalness | instrumentalness | ComParE | psySharpness_sma_stddev | sharpness_std |
| Spotify | valence | valence | ComParE | zcr_sma_amean | zcr[2] |
| Spotify | tempo | tempo | ComParE | zcr_sma_stddev | zcr_std |

[1] The tag with highest confidence is used to represent the genre of music.

[2] zcr, zero-cross rate, is a commonly used feature in music information retrieval.

corresponding *weather* features are represented with conditions (sunny, rainy, or cloudy), temperature, humidity, and pressure.

- **Biological signals and Activities**: Biological signals (i.e., heart rate) and activities (i.e., steps, activity intensity, and activity kind) are all collected at 1Hz, and the missing values are filled by interpolation. Then, 30 minutes of data before each music listening task serves as sequential features.
- **Mood**: Each mood label is scaled to the range of [-1,1] in Arousal and Valence dimensions, respectively. Hereafter, we define the mood recorded before music listening as $\mathbf{mood}_{pre}$ and the mood after as $\mathbf{mood}_{post}$.

It is worth mentioning that, although heart rate variability (HRV) is widely used for emotion recognition [3, 33, 54], it is not available in our experiment, as the record frequency is too low. Moreover, heart rates contain some missing data. A PPG sensor with LED as the light source is used for heart rate detection in MiBand3[8].In the field study, as participants wear the bracelet in daily life doing various activities, poor contact and motion artifacts happen from time to time [9]. However, experiments with low-frequency while low-cost wearable devices are of great value, as they are especially portable and widely used in everyday life. More discussions about the devices are shown in Section 7.2.

In the end, music recommendation records without locations in an hour or with missing biological signals in more than 15 in 30 minutes before music listening are excluded. Finally, 897 records from 30 participants remain, with 29.9 records (SD=8.44) per participant on average, ranging from 11 to 49 records.

### 4.5 Statistics of Participant Preference and Data Collection

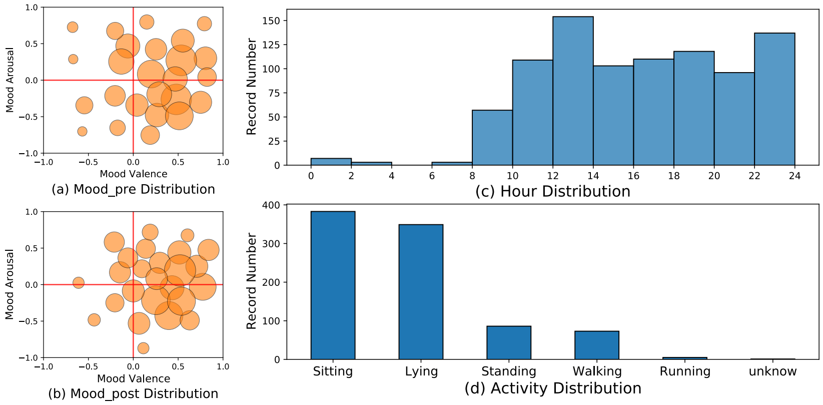
In the pre-experiment interview, we required participants to recall their preferred music genres, and the records revealed diverse music preferences: 76.7% of participants mentioned that they like *pop*, 66.7% mentioned *folk*, 56.7% mentioned *light*, 30.0% mentioned *classical. Rock*, *hip-hop*, *jazz*, *blues*, and *dance* were also mentioned by several (6.7%-16.7%) participants.

Table 2 and Figure 7 provide distributions of music preference (ratings) and various contexts. Most music records gain ratings 2 (20.96%), 3 (32.78%), and 4 (33.22%), while extreme ratings, 1 and 5, are less common, which sum up to 13.04%. Most of the mood records are positive (76.25% in $\text{mood}_{pre}$, and 80.60% in $\text{mood}_{post}$), and distribution of $\text{mood}_{post}$ is slightly more positive than $\text{mood}_{pre}$, as shown in Figure 7(a)/(b). As for context, most

[8] https://www.mi.com/shouhuan3/specs

Table 2. Distribution of ratings on all recommended music tracks in the field study.

| **Rating** | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **No. (%)** | 29 (3.23%) | 188 (20.96%) | 294 (32.78%) | 298 (33.22%) | 88 (9.81%) |



(a) Mood_pre Distribution

(c) Hour Distribution

(b) Mood_post Distribution

(d) Activity Distribution

Fig. 7. Distributions of context information in the field study, including mood, time, and activity. Mood distribution is represented by K-means clusters (cluster number = 25) of records on the two-dimensional mood map. In Figure(a) and (b), the circle center is the clustering center, and the radius represents the number of records in each cluster.

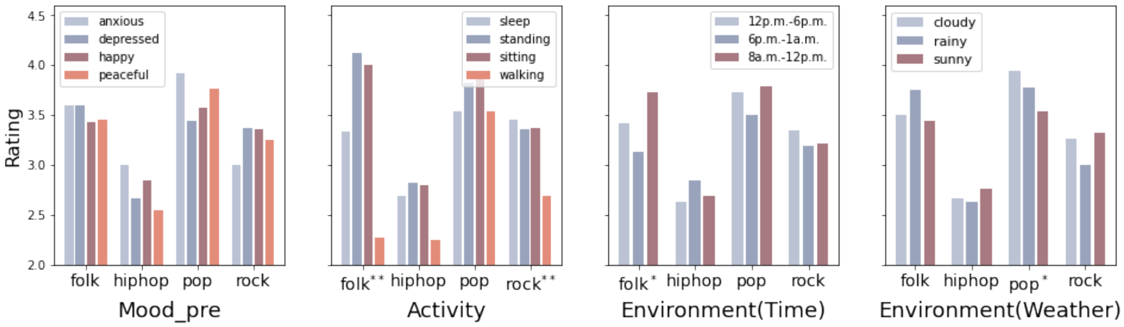


Fig. 8. Average ratings of four typical genres when participants are in different context.

listening tasks were finished between 8 in the morning and 12 in the midnight (Figure 7c) when participants were sitting or lying (Figure 7d).

## 5 FINDINGS ON FIELD STUDY ABOUT UBIQUITOUS MUSIC RECOMMENDATION

In this section, we provide statistical analyses of music preferences, mood, and bracelet-based contexts in the field study to explore the influence of contexts on music preferences in an uncontrolled environment.

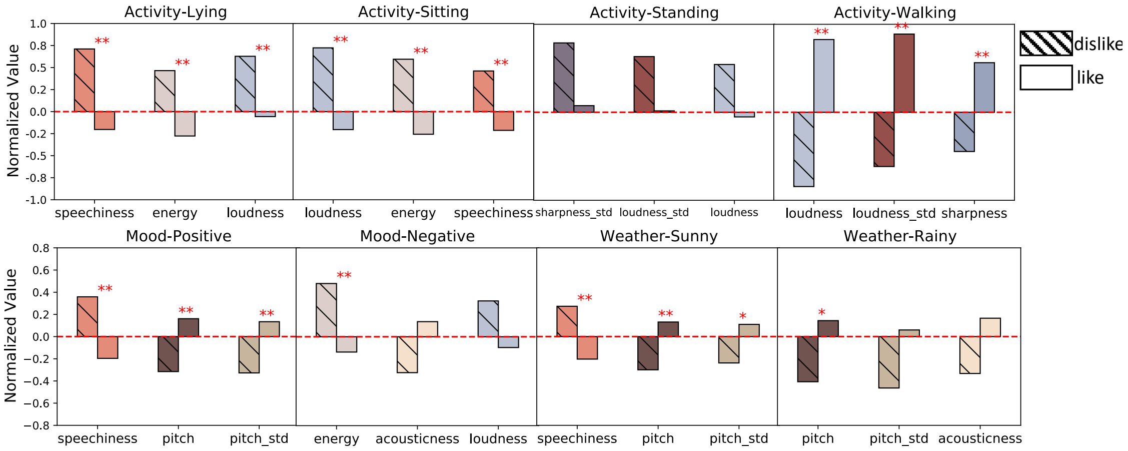


Fig. 9. Three audio features with largest distances between *liked* and *disliked* music in different contexts. $*/**$ indicate significant differences between *like* and *dislike* at $p < 0.05/0.01$ of t-test, respectively, with Holm correction on all audio features in the same situation.

## 5.1 Ubiquitous Context Influence on Music Genre Preference

We start by discussing how the contexts influence preferences for different music genres. Figure 8 presents the average ratings in different contexts of the top-four main genres in the 1000-track music dataset, i.e., *country*, *hip-hop*, *pop*, and *rock*. Moods are divided into four categories according to the quadrant they belong to in the Thayer model.

Figure 8 shows that participants have different preferences for music genres in different contexts. Consistent with analyses of survey on preferred music genres in different moods in Figure 3, Figure 8(a) indicates that **mood** has an influence on music preference. For instance, ratings for folk music are higher when users are anxious or depressed than when they are happy or peaceful. However, mood-related rating differences are not significant in our dataset mainly due to sparse data points in anxious and depressed moods. In different **activities** (Figure 8b), ratings for the same genre are also different, and the differences are significant for folk and rock music. For both folk and rock, participants tend to give low ratings when walking and high ratings when standing, sleeping, and sitting. As for **environment** (time in Figure 8c and weather in Figure 8d), participants tend to favor folk more in the morning than in the afternoon, and rating for pop music is significantly distinctive in different weather conditions. Moreover, no context causes consistent low or high ratings in general. As for genres, although hip-hop music usually gets low ratings, its ratings still vary with contexts.

The above observations suggest that music genre preference is changing with mood, activity, and environment. Thus, it is valuable to conduct ubiquitous personalized recommendation to recommend different music genres under diverse contexts.

## 5.2 Ubiquitous Context Influence on Audio Features Preference

Secondly, we analyze music preferences under different contexts and audio features in Table 1. In each context, we calculate the average value of audio features of the tracks in *like* records (rating>3) and in *dislike* records (rating<3), respectively. Three audio features with the largest distances between *like* and *dislike* are shown in Figure 9. To better compare different features, we operate normalization on each audio feature, and perform an independent t-test for each audio feature in every pair of *like* and *dislike* tracks. To counteract the problem of multiple

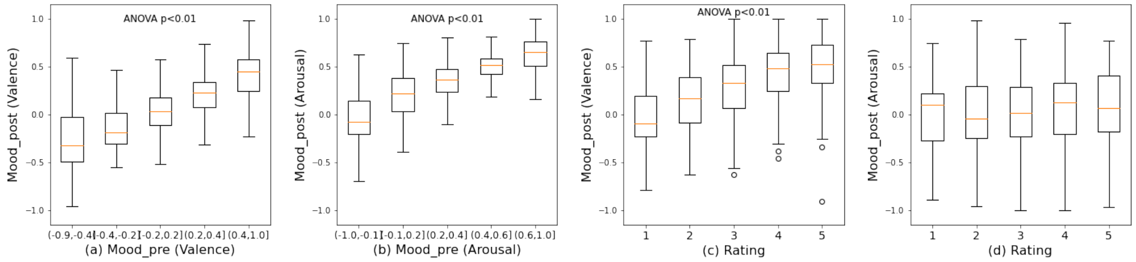


Fig. 10. Distribution of $\text{mood}_{post}$ (in terms of valence and arousal) in five equal-sized bins of $\text{mood}_{pre}$ (Figure a and b), and in five rating levels (Figure c and d).

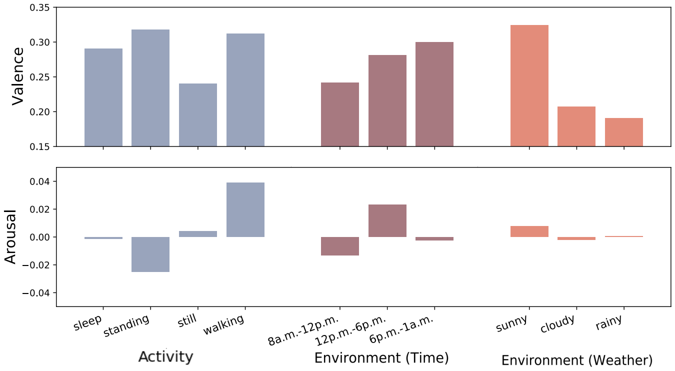


Fig. 11. Average $\text{mood}_{pre}$ (in terms of valence and arousal) in different contexts. The differences of valence under different weather (Right in the first row) and different activities (Left in the first row) are significant, with $p < 0.01$ and $p < 0.05$ respectively with unidirectional ANOVA test.

comparisons, we correct the t-test p-value with the Holm-Bonferroni method on all features in a context situation [31]. We choose the Holm method because it is widely used in hypothesis testing and offers a simple and powerful test to control the family-wise error rate. $*$ and $**$ indicate $p < 0.05$ and $p < 0.01$ after correction, respectively.

Figure 9 indicates that music preference is related to different audio features in distinct situations, i.e., three audio features with the largest distances are different between contexts. Up to nine different audio features appear in the top-three audio features in eight situations. For instance, *speechiness* between *like* and *dislike* music is most disparate in *lying*, *positive mood*, and *sunny day*, while *loudness* is most disparate for *sitting* and *walking*. Meanwhile, the same audio feature may lead to opposite preferences in different situations. For example, *loud* music is preferred when participants were *walking*, whereas low *loudness* is favored when *lying* or *sitting*.

Thus, music preferences, in terms of audio features, are distinct in different contexts. This finding further motivates us to perform ubiquitous music recommendation considering interplay of contexts and audio features.

## 5.3 Relationship between Mood, Music Preference, and Bracelet Data

Survey results in Section 3 illustrate that mood is an essential factor influencing music preference, whereas mood is difficult to collect in practice. To further explore whether and how we should consider mood in music

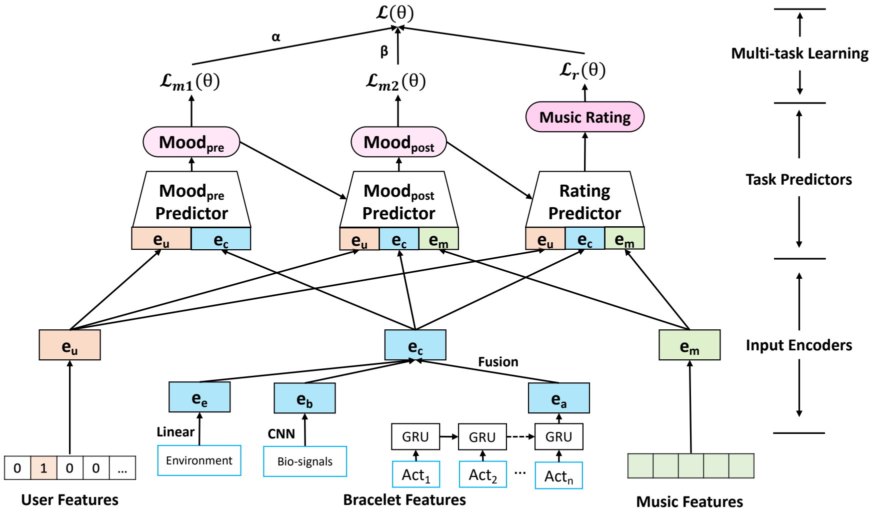


Fig. 12. An overview of the Multi-task Ubiquitous Music Recommendation model (MUMR).

recommendation, we try to analyze the relationship between mood and music preference, as well as mood and contexts.

From participants' feedback, we analyze the distribution of mood$_{post}$ in five-bin mood$_{pre}$ and five-level ratings, respectively, as shown in Figure 10. It demonstrates that mood$_{post}$ and mood$_{pre}$ are significantly related in both valence and arousal dimensions (with Pearson's $r$ = 0.674 and 0.625). Meanwhile, there is a significant proportional relationship between mood valence and preference ratings. Therefore, music preference is closely related to mood, especially valence of mood$_{post}$. Moreover, previous research has manifested it possible to predict mood with data collected by low-cost portable devices [41]. Here we consider average mood$_{pre}$ under different situations in Figure 11. The valence of mood$_{pre}$ is significantly different between different activities and weather conditions, while the differences in arousal dimension is not apparent. It indicates that mood$_{pre}$ is at least partially predictable with bracelet contexts.

In conclusion, it is necessary to consider mood in ubiquitous music recommendation. Meanwhile, mood can be inferred from contexts.

In this section, we deeply analyze the influence of real-world contexts on music preference and explore the relationship between mood and music preference. The findings reveal that contexts and mood both significantly influence music preference. Therefore, in the ubiquitous personalized music recommendation model in Section 6, contexts are used as input for preference prediction, and mood serves as an auxiliary prediction task. On the other hand, the above analysis is focused on the average effects on all participants. However, due to participants' distinct inherent preferences, influences of context and mood may be different at the individual level. Therefore, we further discuss the participant heterogeneity in Appendix B.

## 6 PERSONALIZED UBIQUITOUS MUSIC RECOMMENDATION MODEL AND RESULTS

Since we have found a close relationship between bracelet contexts, mood, and music preference in the field study, we further propose a model to predict music preference with the help of contexts and mood.

## 6.1 Multi-task Ubiquitous Music Recommendation Model (MUMR)

Firstly, we introduce the design of the Multi-task Ubiquitous Music Recommendation model (MUMR), a multi-task optimization method to perform ubiquitous personalized music preference prediction. An overview of the model is shown in Figure 12, which consists of input encoders, task predictors, and multi-task learning modules. In this model, the main task is to predict music preference ratings for a user-music pair. Bracelet contexts are used as input for predictors, and mood prediction is incorporated as an auxiliary task.

*6.1.1 Input Encoder.* Three information sources are fed into MUMR: *User*, *Music*, and *Bracelet*. With $N$ users in dataset, *User* is represented by one-hot embedding $\vec{u} \in R^{N \times 1}$ of user IDs, which contributes to the personalization of the model [13, 26, 68]. Twenty audio features in Table 1 denotes features $\vec{m} \in R^{M \times 1}$ for *Music*, where *key* and *genre* are one-hot encoded, and the others are normalized. *User* and *Music* are embeded into vectors $\vec{e}_u \in R^{u_n \times 1}$ and $\vec{e}_m \in R^{m_n \times 1}$. As for *Bracelet*, *environment* contexts are encoded into $\vec{e}_e$ with a linear layer. A sequence of *biological signals* (i.e., heart rate) is encoded as $\vec{e}_b$ with CNN. The *activity* sequence is encoded with GRU model [16], and the last hidden state $\vec{h}_n$ is used as the activity embedding $\vec{e}_a$. In the end, $\vec{e}_e$, $\vec{e}_b$, and $\vec{e}_a$ are fused to represent *Bracelet*: $\vec{e}_c \in R^{c_n \times 1}$.

*6.1.2 Task Predictors.* We perform prediction for Mood$_{pre}$, Mood$_{post}$, and Rating (i.e., music preference) with three separate predictors. Motivated by the relationship between mood and ratings in Figure 10, the output of Mood$_{pre}$ predictor is utilized as input for Mood$_{post}$, and the output of Mood$_{post}$ predictor is fed into the Rating Predictor. All three predictors are implemented with Multi-Layer Perception (MLP),

$$\vec{m}'_{pre} = \sigma(\phi_{m1}(\phi_{an}(...\phi_{a2}(\phi_{a1}(\vec{e}_u \oplus \vec{e}_c))...))) \tag{1}$$

$$\vec{m}'_{post} = \sigma(\phi_{m2}(\phi_{bn}(...\phi_{b2}(\phi_{b1}(\vec{e}_u \oplus \vec{e}_c \oplus \vec{e}_m \oplus \vec{m}'_{pre}))...))) \tag{2}$$

$$r' = \sigma(\phi_r(\phi_{rn}(...\phi_{r2}(\phi_{r1}(\vec{e}_u \oplus \vec{e}_c \oplus \vec{e}_m \oplus \vec{m}'_{post}))...))) \tag{3}$$

Where $\phi_{m1}$, $\phi_{m2}$, and $\phi_r$ are linear layers with output dimensions two (for mood valence and arousal, respectively), two, and one. $\sigma$ denotes sigmoid function. $\phi_{ai}$, $\phi_{bi}$, and $\phi_{ri}$ are all linear layers: $\phi_k(\vec{x}) = relu(\mathbf{W_k}\vec{X} + \vec{b_k})$, and there are $n$ layers in total. $\oplus$ indicates concat operation of two vectors. $\vec{m}'_{pre}$, $\vec{m}'_{post}$, and $r'$ are predicted values for Mood$_{pre}$, Mood$_{post}$, and Rating, respectively.

*6.1.3 Multi-task Learning.* Multi-task learning (MTL) is a paradigm to jointly train different but correlated tasks to obtain better performance for each task [21]. Considering the tight connections between mood and music preference in the previous analyses, we utilize mood prediction as an auxiliary task for rating prediction to provide cross-task information. To be specific, we perform multi-task optimization of mood and rating, with Mean Squared Error (MSE) serving as loss for all tasks,

$$\mathcal{L}_{m1}(\theta) = \frac{1}{K}\sum_{k=1}^{K} ||(\vec{m}_{k},_{pre} - \vec{m}_{k},'_{pre})||^2 \tag{4}$$

$$\mathcal{L}_{m2}(\theta) = \frac{1}{K}\sum_{k=1}^{K} ||(\vec{m}_{k},_{post} - \vec{m}_{k},'_{post})||^2 \tag{5}$$

$$\mathcal{L}_{r}(\theta) = \frac{1}{K}\sum_{k=1}^{K} (r - r')^2 \tag{6}$$

Where $\vec{m}_{kpre}$, $\vec{m}_{kpost}$, and $r$ are true labels, $K$ is the training set size, $k$ indicates the k$th$ sample, $||\vec{X}||$ is the norm of vector $\vec{X}$, and $\theta$ represents model parameters. The final loss is the weighted sum of three tasks:

$$\mathcal{L}(\theta) = \mathcal{L}_r(\theta) + \alpha \cdot \mathcal{L}_{m1}(\theta) + \beta \cdot \mathcal{L}_{m2}(\theta) \tag{7}$$

$\alpha$ and $\beta$ are both hyper-parameters to control weights for the auxiliary tasks, i.e., mood predictions.

## 6.2 Experimental Settings

In this section, we elaborate the evaluation metrics, baselines, and experimental settings.

*6.2.1 Dataset.* We conduct experiments on the dataset from field study in Section 4.4, including 30 participants, 105 music tracks, and 897 interaction records. The lengths of user vector $\vec{u}$ and music features $\vec{m}$ are $N = 30$ and $M = 43$, respectively. As for contexts, the length of environment features is 11 (3 for one-hot time, 2 for location, and 6 for weather). The biological signal and activity are both sequential data with lengths of 30. Bio-signal is one-dimensional (heart rate), and activity is 7-dimensional (1 for step, 1 for activity intensity, and 5 for one-hot activity type).

*6.2.2 Evaluation Protocols.* We perform 10-fold cross-validation for evaluation. At each fold, we randomly take 90% of interaction records for training and the remaining for testing. As for labels, moods in each dimension are scaled into $[-1, 1]$. And music ratings are re-split into three levels in model training and testing to balance distribution of ratings: 1 for rating<3, 2 for rating=3, and 3 for rating>3. The performance is estimated by Mean Squared Error (MSE) for both rating and mood predictions.

*6.2.3 Baselines.* We compare MUMR with the following two methods:

- **GBRT**: Gradient Boosted Regression Trees. A conventional regression model with ensemble learning. User id, music features, and bracelet contexts are concatenated as input for **GBRT**.
- **Wide & Deep**[13]: A personalized recommender that jointly learns *Wide* linear models and *Deep* neural networks. Following the original work, the *Wide* side takes user and music as input, and the *Deep* side takes user, music, and bracelet contexts as input.

Notice that our primary purpose is to detect the usability of bracelet contexts for music recommendation. And due to the limited dataset size, sophisticated deep models are impractical for training. Therefore, further explorations on model structures and baseline comparisons is left for future work.

*6.2.4 Parameter Settings.* We tune the hyper-parameters according to the best performance of rating predictions.

For MUMR, a linear layer with ReLU activation is adopted for user and music encoders, with $u_n = m_n = 16$ as embedding size. Encoder dimensions for each type of bracelet context are tuned separately, and the output are concatenated to fuse $\vec{e}_c$ with $c_n = 24$. MLP with a hidden layer of 128 neurons is adopted for all predictors. Sigmoid function is used for the prediction layer. And the loss is calculated with $\alpha = 0.05$ and $\beta = 0.2$. For the baselines, Wide & Deep contains two hidden layers at the Deep side with 64 neurons and 16 neurons, respectively. GBRT adopts 200 estimators with a maximum depth of 5 for all trees.

Two neural models (i.e., MUMR and Wide & Deep), are trained with Adam [36]. The learning rate is 0.002, and early-stopping is conducted with patience of 20 epochs, with maximal epochs of 300. To prevent models from overfitting, we use the $\ell_2$-regularization with weight 2e-02.

## 6.3 Performance

*6.3.1 Overall Performance.* The overall performances of MUMR and two baseline methods are shown in the first and second parts of Table 3. We have the following observations:

Table 3. Overall performance of our proposed approach, MUMR, and baseline methods in terms of Mean Squared Error (MSE), where the lower score indicates better performance. $/m$ indicates dropping mood, and $/c$ indicates not considering bracelet context. **Bold** fonts indicate the best results in the group. Two-side t-test is conducted for 10-fold cross validation. $^*/^{**}$ indicate p-value<0.05/0.01 when comparing with the best baseline in the same group. $^+/^{++}$ indicate p-value<0.05/0.01 when comparing with MUMR using *activity, env, bio*. Underline indicates p-value<0.05 when comparing with the same method in the no context group.

| **Group** | **Model** | **Context Features** | **Task1: Rating (MSE)** | **Task2: Mood$_{pre}$ (MSE)** | | **Task3: Mood$_{post}$ (MSE)** | |
|---|---|---|---|---|---|---|---|
| | | | Preference | Valence | Arousal | Valence | Arousal |
| No Context | GBRT/c | - | 0.511 | - | - | - | - |
| | WD/c | - | 0.482 | - | - | - | - |
| | MUMR/c | - | **0.474** | **0.144** | **0.171** | **0.100** | **0.121** |
| With Context | GBRT | activity,env,bio | 0.441 | - | - | - | - |
| | WD | activity,env,bio | 0.402 | - | - | - | - |
| | MUMR | activity,env,bio | 0.350$^*$ | **0.126** | **0.161** | **0.106** | **0.118** |
| Ablation Study | MUMR/m/c | - | 0.480$^{++}$ | - | - | - | - |
| | MUMR/m | activity,env,bio | 0.371$^+$ | - | - | - | - |
| | MUMR | env,bio | 0.428$^{++}$ | 0.135 | 0.178 | 0.109 | **0.125** |
| | MUMR | activity,bio | 0.381$^+$ | 0.145 | 0.180 | **0.106** | 0.128 |
| | MUMR | activity,env | **0.365** | **0.136** | **0.172** | 0.107 | 0.129 |

Firstly, comparing the results without and with context (i.e., the first and second group), we find that adding contexts will promote rating prediction performance significantly for all three methods. It illustrates that music preference can be better modeled with bracelet collected contexts, even by simply concatenating the context as input in GBRT and Wide & Deep. And more improvement is achieved with the input encoders in MUMR, resulting in 26.16% decay of rating prediction MSE (from 0.474 to 0.350).

Secondly, MUMR achieves the best performances in both settings. MUMR/c has similar performance with Wide & Deep because MUMR/c is similar to the Deep side of Wide & Deep adding joint learning of mood. With contexts, MUMR performs significantly better than Wide & Deep, which indicates the importance for a proper method to encode the bracelet contexts.

To better illustrate the effectiveness of contexts, we show the confusion matrix of MUMR/c and MUMR on the rating prediction task in Figure 13. In the triple classification task, the accuracy of MUMR/c is 50.2%, and the accuracy of MUMR is 62.9%. For each class, with MUMR/c, the f1-scores of predicting Label 1, Label 2, and Label 3 are 0.456, 0.491, 0.543, respectively. For MUMR, the f1-scores are 0.687, 0.521, and 0.695. And the recalls are 36.0%, 69.2%, and 43.8% for MUMR/c, and 64.0%, 58.2%, and 65.9% for MUMR. Compared with MUMR/c, MUMR gains higher f1-scores for all three classes. The recall at Label 2 is lower for MUMR, because MUMR generates fewer average ratings (i.e., Label 2). Moreover, MUMR is more accurate at predicting *dislike* and *like* status (i.e., Label 1 and Label 3), which indicates that MUMR performs more useful music recommendation for users.

When considering the mood prediction of MUMR and MUMR/c in Table 3, we find that contexts help predict Mood$_{pre}$, while do not improve Mood$_{post}$ prediction accuracy. It is because MUMR/c predicts Mood$_{pre}$ with user information only, adding context features in MUMR is helpful. However, prediction for Mood$_{post}$ relies on Mood$_{pre}$, user, and music features in MUMR/c. Hence, in MUMR, adding context features is less influential for Mood$_{post}$ prediction.

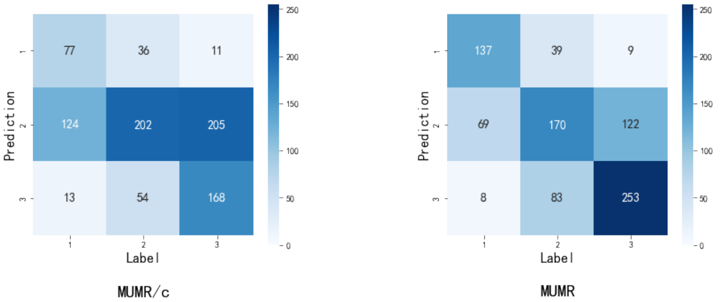


Fig. 13. Confusion matrix of MUMR/c (left) and MUMR (right) on rating prediction task.

*6.3.2 Ablation Study.* To clarify the contribution of contexts, we compare MUMR with a series of baselines that consider partial contexts and mood information, as shown in the third part of Table 3.

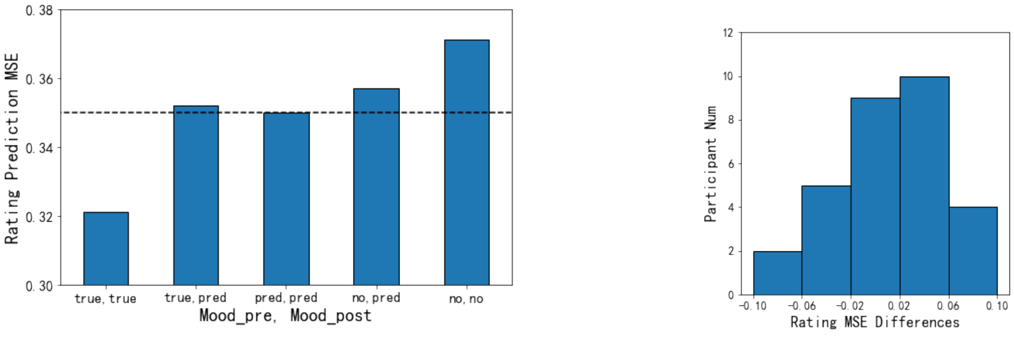
Compared with MUMR/m/c and MUMR/c, MUMR/m and MUMR achieve significant improvement, respectively. It illustrates that bracelet contexts are essential for rating prediction. Meanwhile, the mood contributes to slight but significant improvement between MUMR/m and MUMR, which demonstrates the effectiveness of using mood as a latent factor. Detailed discussions on utilization of mood will be shown in Section 6.4.1. Furthermore, comparing three types of bracelet contexts, we find that dropping activity leads to the largest performance decrease, followed by environment, and then biological signals. For one thing, ongoing activity has a significant influence on music preference, as analyzed in Figure 8. For another, activity detection by low-cost smart bracelets is relatively accurate with current devices. Environment also helps predict music preference with significant performance improvement, same as in survey results in Figure 1. The biological signals, i.e., heart rate sequence, show no significant improvement on rating prediction. It is mainly due to that heart rate sampling frequency is too low, and missing value is also a problem. The problem has been mentioned in Section 4.4, and we further discuss the usage of heart rate signals in Section 7.2.

In conclusion, both the bracelet context and mood factors are effective in MUMR. As for different contexts, activity and environment significantly contribute to music preference prediction, while biological signals by low-cost bracelets gain limited improvement. Portable devices with higher sampling rates may improve the performance of bio-signals, which is left for the future works.

## 6.4 Further Analysis on Model Performance

*6.4.1 Mood in Ubiquitous Music Recommendation.* As discussed in the analysis of questionnaires and user study, mood is an important factor in music preference prediction. Nevertheless, mood status is difficult to collect in practice. In this section, we further analyze the usage of mood in MUMR.

Firstly, we try different representations of Mood$_{pre}$ and Mood$_{post}$ as inputs for Mood$_{post}$ Predictor and Rating Predictor, respectively. As shown in Figure 14a, predictors are trained with true mood labels, predicted mood, or without mood. The results demonstrate improvement when true labels for Mood$_{post}$ are used as input (i.e., true, true), which is in line with the significant correlation between Mood$_{post}$ and Rating (Figure 10). Nevertheless, Mood$_{post}$ is unavailable in practice. If true labels for Mood$_{pre}$ are applied only (i.e., true, pred), the preference prediction performance is similar to that of predicted moods for both (i.e., pred, pred), which indicates that Mood$_{pre}$ is less influential on rating. Furthermore, the performance will decay if one or no mood predictors are implemented in MUMR (no, pred and no, no). It illustrates that the auxiliary mood prediction task provides supplementary information for rating prediction task. It is also worth noticing that the performance with no



(a)

(b)

Fig. 14. (a) MUMR rating prediction MSE on different settings of mood, where *true,pred* is using true $\text{Mood}_{pre}$ labels as the input of *$Mood_{post}$ predictor* and using predicted $\text{Mood}_{post}$ as the input of *Rating predictor*, and so on. (b) Distribution of distance between rating prediction MSE without and with mood for each participant.

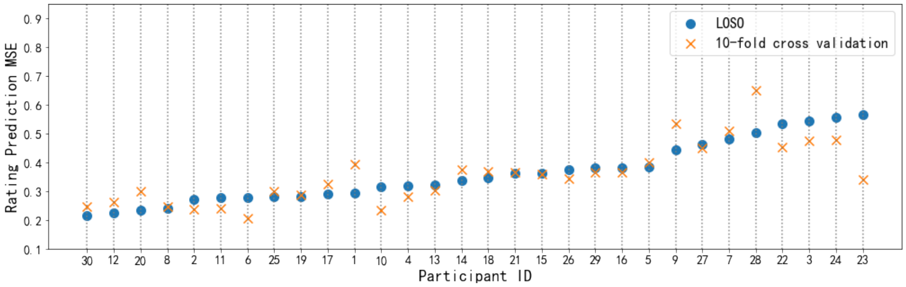


Fig. 15. Comparison of Leave-One-Subject-Out evaluation and 10-fold cross validation on each user in terms of rating prediction MSE, where lower MSE indicates better performances.

mood predictor is fairly good (MSE=0.371), so accurate music rating prediction can be achieved with bracelet contexts even without additional mood labels.

Then we discuss another condition assuming some participants do not record their moods. For each participant $u$, we conduct MUMR with multi-task loss for the other participants but rating loss alone for $u$, which results in rating prediction MSE $\mathcal{L}'_r(u)$. We compare $\mathcal{L}'_r(u)$ with $\mathcal{L}_r(u)$ (MSE on $u$ with full information) on samples of $u$. The distribution of the differences $d = \mathcal{L}'_r(u) - \mathcal{L}_r(u)$ on 30 participants is shown in Figure 14b. Participants' own mood labels help promote rating prediction for 14 participants while hurt prediction for 7 participants. And models for 9 participants achieve similar results under two conditions ($d \in [-0.02, 0.02)$). For most participants, considering personalized moods at least does not hurt preference understanding. On the individual level, mood is not always helpful for music preference prediction, which may result from personalized sensitivity to mood status.

Therefore, the multi-task strategy is generally helpful for music preference modeling, while MUMR also performs well for users without mood records.

*6.4.2 Leave-One-Subject-Out Evaluation.* Cold-start user, i.e., user with no historical feedback information, is a common and important problem in recommender systems, requiring flexibility of models [39]. So we propose

Leave-One-Subject-Out (LOSO) evaluation for MUMR, keeping data for one participant as the test set and the others as training set at one time. By LOSO evaluation, we validate whether MUMR can deal with the cold-start user problem.

Performances on each of 30 participants with 10-fold cross-validation and LOSO evaluation are shown in Figure 15. It illustrates that the performance of LOSO evaluation is not always worse than cross-validation. To be specific, 10 participants have similar performance under two evaluation methods with MSE differences less than 0.02, and better performance is achieved for 10 participants if he/she is treated as new users. It may be because these participants have similar music preferences to the whole dataset. However, LOSO evaluation has worse performance on 10 participants, especially for Participant 1, 28, 9, and 20 (with MSE differences lower than -0.05). These participants may have particular preferences for music hard to model with other participants' records. The largest MSE in LOSO is 0.57, while it is 0.65 in cross-validation. Therefore we conclude that, generally MUMR is effective for new users recommendation.

## 6.5 Real-world Testing for MUMR

*6.5.1 Experiment Procedure.* To further verify the utility of context-aware MUMR, we conduct an experiment in real world with MUMR. The real-world testing follows the same procedures as the field study in Section 4.3.2, except that three recommendation algorithms (*Random*, *Wide&Deep*, and *LightGCN*) are replaced by *Random*, *MUMR/c* (i.e., MUMR without contexts), and *MUMR*. MUMR and MUMR/c are both trained with the best settings in Section 6.2 on the whole field study dataset.

In the 1-week field study, participants are required to wear the same bracelet as in the former field study to record their contexts, and the data will be uploaded to the smartphone by GadgetBridge. Every 1-2 hours in the daytime, participants would be reminded to perform one music listening task on LifeRec. When a participant chooses to listen to music, the context data will be automatically uploaded to the remote server. If it is MUMR's turn as a recommender, all unlistened music tracks in the 1000 selected tracks will be ranked by ratings predicted by MUMR, and one music track with the highest rating will be recommended. The same recommendation strategy is adopted for MUMR/c, except that prediction is made without context features. And Random will recommend one unlistened track randomly. As for feedback, besides the mood and discrete rating shown in Figure 5, we also require users to record continuous preference (from 1 to 100) with a slider added below the rating checkbox.

*6.5.2 Participants.* Since MUMR requires historical recommendation records for training to achieve better personalized recommendation performances, we re-recruit respondents from the 30 participants in the former experiment (which was conducted about 7 months ago). Participants are invited to take the real-world testing if they (1) still use smartphones with Android system, and (2) have no large shift in music preference since previous experiments. Finally, 10 participants agreed to take the real-world testing experiment, and none of them dropped out. Their participant IDs are 9, 11, 13, 14, 15, 18, 19, 21, 26, and 30. Among them, 6 were female, and 4 were male, aged 19-25 (M=22.9, SD=1.73). All participants were informed of the experiment and signed consents again. The financial incentive for participants was the same as in former experiments.

*6.5.3 Results and Discussions.* In the real-world testing, 509 valid records were collected. The participants listened to 50.9 music tracks on average (std = 11.42), ranging from 40 (User 14) to 79 (User 26). Overall and individual average rating and preference are displayed in Table 4 and Figure 16, respectively.

Compared with Random, MUMR/c achieves significant improvement. It demonstrates that our proposed personalized recommenders help predict participants' music preferences more accurately. Besides, MUMR performs better than MUMR/c, significant when comparing the average of participants' mean, which indicates that context features improve recommendation performance. Moreover, there are differences between participants. Six of ten participants, User 9, 13, 14, 19, 21, and 30, prefer music recommended by MUMR to MUMR/c. However, User

Table 4. Average rating (in range of 1-5) and preference (in range of 0-100) in real-world testing. MUMR/c represents MUMR model without context as input. *All Avg.* denotes the mean value of all records. *P. Avg.* denotes the mean value of participants, i.e., mean value is calculated within each participant and then averaged. T-test is conducted between random and MUMR/c, and MUMR/c and MUMR, and */** indicates p-value<0.05/0.01.

| | **All Avg.** | | **P. Avg.** | |
|---|---|---|---|---|
| | **Rating** | **Preference** | **Rating** | **Preference** |
| **Random** | 2.97 | 48.49 | 2.83 | 45.21 |
| **MUMR/c** | 3.35$^{**}$ | 57.42$^{**}$ | 3.36$^{**}$ | 57.46$^{**}$ |
| **MUMR** | **3.53**$^{*}$ | **61.02** | **3.57**$^{*}$ | **61.78**$^{*}$ |

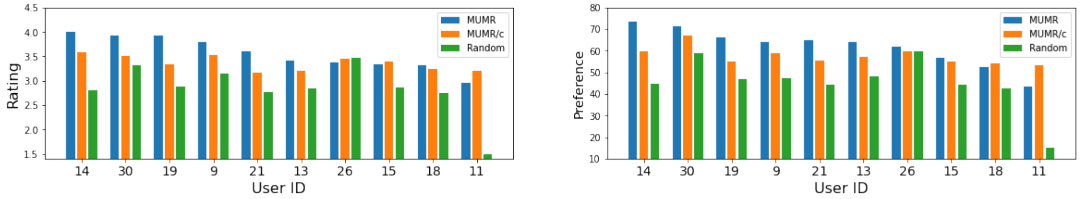


Fig. 16. Average rating and preference of all participants in real-world testing.

15, 18, and 26 show similar preferences for recommendations between MUMR and MUMR/c. These participants may not be sensitive to context information. And User 26 shows receptive to various music. User 11 prefers MUMR/c to MUMR. During the interview with User 11, she mentioned never feeling the recommended music was related to her status during the real-world testing. In conclusion, MUMR shows better performance than MUMR/c in general, while individual usefulness of contexts is correlated with users' characteristics.

As for subjective experience, in the post-experiment interview, we asked the participants whether they were willing to try ubiquitous music recommenders after the experiment. Seven of ten participants would like to use it, three gave neutral feedback, and none are unwilling to try. Therefore, participants showed generally positive attitudes towards ubiquitous recommenders after the real-world testing.

## 7 DISCUSSIONS AND LIMITATIONS

This work shows the potential to conduct ubiquitous personalized music recommendation with low-cost wearable devices. Here we discuss its contribution to the recommendation scenario, the usage of low-cost portable devices, privacy concerns, and some limitations.

### 7.1 Ubiquitous Personalized Music Recommender and its Contribution to Recommendation Scenario

Previous works have demonstrated that music preference is affected by contexts such as mood [23, 75], activity [61], and weather [38] with laboratory experiments and survey analysis, where imagined or induced contexts are used to measure the influence on music preference. Unlike these works, we conduct a field study to collect music preferences under different contexts in real life.

Analyses in Section 5 show differences in music preferences in different contexts including environment, activity, and mood status. And model experiments in Section 6 illustrate that rich contexts from smart bracelet data help promote music recommendation performances. Further analysis demonstrates that it is possible to perform accurate music recommendation for users without mood annotation or even without historical interactions. Real-world testing further illustrates the capability of contexts and our proposed model. Therefore, modeling music preference with bracelet contexts is practical despite the cognitive gap between high-level music preference and bracelet-based data. We can answer the question proposed in Section 1: *It is possible to perform ubiquitous personalized music recommendation with contexts collected by wearable devices.*

Moreover, our approach is not limited to music recommendation. The promising results also encourage exploration in other scenarios.

### 7.2 Adoption of Low-cost Portable Devices

In this work, contexts about activity, environment, and biological signals are collected with low-cost smart bracelets. The ablation study in Section 6.3.2 illustrates that activity has the greatest influence on music preference modeling, environmental features take second place, and biological signals have the least and insignificant influence on the performance.

We speculate that the influence of different information sources is closely related to the characteristics of low-cost portable devices we adopted in the experiment. The smart bracelets can detect steps, activity intensity, and environment information accurately. Meanwhile, activity and environment are steady under the sampling frequency (1Hz) in our experiment. However, detection of biological signals is accompanied by noises and missing values. Although biological signals have been widely used in previous works for mood detection [33, 50, 81], professional devices with a high sampling frequency (usually higher than 10Hz) and abundant types of data (such as GSR and skin temperature) were applied in these works. These high-precision devices can accurately capture subtle changes in heart rate, but they are expensive, not portable, and thus not suitable for daily life. Meanwhile, some previous works showed great associations between heart rate and music listening [44, 53], but they aimed to adjust heart rate with specific music tracks. However, we aim to optimize users' satisfaction, a high-level cognitive process essential for recommenders, which is more difficult to infer with basic heart rate signals. The adoption of low-cost but low-frequency wearable devices for user satisfaction prediction is suitable for daily music application but limits the ability of biological signals to detect personal status.

Nevertheless, it is of great value to attempt low-cost devices since they are close to real-life applications in a foreseeable future than lab-based professional equipment. The encouraging results of our experiments show the possibility of adopting these low-cost wearable devices in both research and application. Meanwhile, we need to point out that better results is expected with high-precision devices. We are looking forward to more attempts in ubiquitous music recommendation with more accurate wearable devices, and we believe the cost of these devices will fall with the development of technology in the future.

### 7.3 Privacy Concern in Personalized Ubiquitous Recommendation

Compared with conventional personalized recommendation methods, the opportunities of constructing personalized ubiquitous recommender systems come with risks for privacy challenges. On the one hand, it is essential to inform users about what data would be collected, how they would be collected, and how they would be used. In our work, each participant was well explained about the information collection and storage strategy with both written informed consent and researchers face-to-face. They also agreed that their data would be used in the following experiments and made public after full anonymity. Furthermore, all collected contexts would be presented to the participant and uploaded to the remote server only after consent. Participants could also choose to delete some of the data before uploading. On the other hand, no personally identifiable information was collected in the experiment, such as surrounding sounds and photos. Among all contexts we collected, the most sensitive information was the participant's GPS. In experiments, we also processed the information to use scaled relative distances instead of specific longitude and latitude records.

However, we must admit that our privacy protection strategy is not enough if the ubiquitous recommendation is used on a large scale. In the future, we will try to deal with the privacy issue from the perspective of recommender models, such as personalized encryption at mobile terminals before uploading, or adoption methods of federated learning [37] in the ubiquitous personalized recommendation.

### 7.4 Limitations

Despite our best efforts, this work still has several limitations. Firstly, the scale of data collection is limited to 30 participants for the field study and 10 for real-world testing. The dataset sample size makes it impossible to conduct ubiquitous music recommendation with most state-of-the-art recommender systems to examine the usefulness of bracelet data more thoroughly. Moreover, participants for the questionnaire survey and user study are primarily students from a public university in China, so the demographic diversity may be insufficient. To remit this problem, we recruit 4 participants from the surrounding community to enrich the diversity. Analyses and experiments on their data show similar results to students, which indicates the stability of our findings to some extent. However, these participants are somehow related to the university (as staff or family members of students). Therefore, it is important to clarify that the findings from our experiments are exploratory, and may not be representative of the music preference of people from other backgrounds. Moreover, some participants are re-recruited for real-world testing, which may lead to bias in testing results. Nevertheless, these limitations should be evaluated under the consideration that we are the first to explore the usage of low-cost bracelet data in ubiquitous recommendation in the uncontrolled environment. Our experimental settings allow us to observe user-music interactions in the wild and collect rich contexts, which we hope to encourage more research on other participant groups.

Secondly, in the recommendation task, we did not take self-reported characteristics (e.g., inherent music preference) into modeling user preference. If these features are considered in the recommender, we may obtain better performance on rating prediction. Nevertheless, it is almost impossible to collect these self-reported features in practice. Therefore, we follow traditional recommender systems and take user IDs and historical preference records to model participants' personalized preferences. Actually, for a ubiquitous recommender system, there are many kinds of context information to consider for improving recommendation accuracy, and more sophisticated methods can be used to better fuse users' physical contexts and online behaviors, which we leave for future work.

Lastly, the overall performance and ablation study shows limited promotion with biological signals. It is mainly related to the low sampling frequency of bracelet data collection and the cognitive gap between low-level bio-signals and the high-level user satisfaction optimization goal. Although there are some high-precision devices for medical or research use (such as Philips' Health Band[9] and Polar chest strap[10]), they are not portable enough to wear daily and are expensive for the general public to use for the ubiquitous recommendation. However, we believe that with the development of portable device techniques, high-quality and high-frequency biological signals can be collected more easily in the future. And bio-signals will make more contributions to ubiquitous recommendation tasks by then.

## 8 CONCLUSION

In this work, we propose the ubiquitous personalized music recommendation with low-cost smart bracelet data. Firstly, a large-scale questionnaire is conducted, which shows that users' music preferences are influenced by their moods, activities, and environments. Especially, we find that subjective mood has a strong influence on music preferences. Since mood can be predicted with environmental and internal status, it is treated as a latent factor in our experiments. Then we perform a one-week user study in the real world with 30 participants to collect context information using smart bracelets and participants' rating preference and mood for recommended music tracks

[9] https://www.usa.philips.com/healthcare/product/HC422210064081/philips-health-band
[10] https://www.polar.com/us-en/sensors/h10-heart-rate-sensor/

in daily life. The user study results show a significant relationship between music preference, user mood, and bracelet data. Based on the data collection, we propose a novel Multi-task Ubiquitous Music Recommendation model (MUMR) to predict users' music preference with bracelet contexts as input and mood prediction as an auxiliary task. Experiments on MUMR demonstrate significant improvement with context information and multi-task training. Further analysis shows that the model also performs well for users without mood labels, which allows practical applications. And real-world testing on 10 participants also illustrates the effectiveness of MUMR. The promising results of data analysis and experiments illustrate that ubiquitous personalized music recommendation is possible with real-world contexts by low-cost smart bracelets. Our work provides a new direction for conducting music recommendation in daily life. Moreover, our approach is not limited to the music scenario, and the proposed model is extensible for other kinds of real-world contexts and recommendation tasks, which we leave as future work.

## ACKNOWLEDGMENTS

We sincerely thank our anonymous reviewers for their insightful feedback. We also would like to thank all participants who took part in our survey and user study. This work is supported by the Natural Science Foundation of China (Grant No. U21B2026) and Tsinghua University Guoqiang Research Institute.

## REFERENCES

[1] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information systems (TOIS)* 23, 1 (2005), 103–145.

[2] Pedro Álvarez, Francisco Javier Zarazaga-Soria, and Sandra Baldassarri. 2020. Mobile music recommendations for runners based on location and emotions: The DJ-Running system. *Pervasive Mob. Comput.* 67 (2020), 101242. https://doi.org/10.1016/j.pmcj.2020.101242

[3] Bradley M Appelhans and Linda J Luecken. 2006. Heart rate variability as an index of regulated emotional responding. *Review of general psychology* 10, 3 (2006), 229–240.

[4] Willian Assuncao, Lara S. G. Piccolo, and Luciana A. M. Zaina. 2022. Considering emotions and contextual factors in music recommendation: a systematic literature review. *Multim. Tools Appl.* 81, 6 (2022), 8367–8407. https://doi.org/10.1007/s11042-022-12110-z

[5] Abigail Bartolome, Sahaj Shah, and Temiloluwa Prioleau. 2021. GlucoMine: A Case for Improving the Use of Wearable Device Data in Diabetes Management. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–24.

[6] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. [n. d.]. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)* (2011).

[7] Hafiz Syed Muhammad Bilal, Wajahat Ali Khan, and Sungyoung Lee. 2017. Unhealthy dietary behavior based user life-log monitoring for wellness services. In *International Conference on Smart Homes and Health Telematics*. Springer, 73–84.

[8] Ian Brace. 2018. *Questionnaire design: How to plan, structure and write survey material for effective market research*. Kogan Page Publishers.

[9] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. 2018. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics* 4, 4 (2018), 195.

[10] Liqiong Chang, Jiaqi Lu, Ju Wang, Xiaojiang Chen, Dingyi Fang, Zhanyong Tang, Petteri Nurmi, and Zheng Wang. 2018. SleepGuard: Capturing rich sleep information using smartwatch sensing data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–34.

[11] Jinpeng Chen, Pinguang Ying, and Ming Zou. 2019. Improving music recommendation by incorporating social influence. *Multimedia Tools and Applications* 78, 3 (2019), 2667–2687.

[12] Toly Chen and Yu Hsuan Chuang. 2018. Fuzzy and nonlinear programming approach for optimizing the performance of ubiquitous hotel recommendation. *Journal of Ambient Intelligence and Humanized Computing* 9, 2 (2018), 275–284.

[13] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. [n. d.]. Wide & Deep Learning for Recommender Systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems* (Boston MA USA, 2016-09-15). ACM, 7–10. https://doi.org/10.1145/2988450.2988454

[14] Zhiyong Cheng and Jialie Shen. 2016. On effective location-aware music recommendation. *ACM Transactions on Information Systems (TOIS)* 34, 2 (2016), 1–32.

[15] Ming-Chuan Chiu and Li-Wei Ko. 2017. Develop a personalized intelligent music selection system based on heart rate variability and machine learning. *Multimedia Tools and Applications* 76, 14 (2017), 15607–15639.

[16] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[17] Ian Cross. 2001. Music, cognition, culture, and evolution. *Annals of the New York Academy of sciences* 930, 1 (2001), 28–42.

[18] Minh Son Dao, Duc Tien Dang Nguyen, Asem Kasem, et al. 2018. HealthyClassroom-a proof-of-concept study for discovering students' daily moods and classroom emotions to enhance a learning-teaching process using heterogeneous sensors. (2018).

[19] Toon De Pessemier, Simon Dooms, and Luc Martens. 2014. Context-aware recommendations through context and activity recognition in a mobile environment. *Multimedia Tools and Applications* 72, 3 (2014), 2925–2948.

[20] Greg T Elliott and Bill Tomlinson. 2006. PersonalSoundtrack: context-aware playlists that adapt to user pace. In *CHI'06 extended abstracts on Human factors in computing systems*. 736–741.

[21] An Evgeniou and Massimiliano Pontil. 2007. Multi-task feature learning. *Advances in neural information processing systems* 19 (2007), 41.

[22] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The World Wide Web Conference*. 417–426.

[23] Ronald S Friedman, Elana Gordis, and Jens Förster. 2012. Re-exploring the influence of sad mood on music preference. *Media Psychology* 15, 3 (2012), 249–266.

[24] Damianos Gavalas and Michael Kenteris. 2011. A web-based pervasive recommendation system for mobile tourist guides. *Personal and Ubiquitous Computing* 15, 7 (2011), 759–770.

[25] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–28.

[26] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).

[27] Casper Hansen, Christian Hansen, Lucas Maystre, Rishabh Mehrotra, Brian Brost, Federico Tomasi, and Mounia Lalmas. 2020. Contextual and sequential user embeddings for large-scale music recommendation. In *Fourteenth ACM Conference on Recommender Systems*. 53–62.

[28] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 639–648. https://doi.org/10.1145/3397271.3401063

[29] Jennifer Healey, Lama Nachman, Sushmita Subramanian, Junaith Shahabdeen, and Margaret Morris. 2010. Out of the lab and into the fray: Towards modeling emotion in everyday life. In *International Conference on Pervasive Computing*. Springer, 156–173.

[30] Ramón Hermoso, Sergio Ilarri, Raquel Trillo, and María del Carmen Rodríguez-Hernández. 2015. Push-based recommendations in mobile computing using a multi-layer contextual approach. In *Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia*. 149–158.

[31] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[32] Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2016. Multi-task learning for predicting health, stress, and happiness. In *NIPS Workshop on Machine Learning for Healthcare*.

[33] Eiman Kanjo, Eman MG Younis, and Chee Siang Ang. 2019. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion* 49 (2019), 46–56.

[34] Ken Kawamoto, Takeshi Tanaka, and Hiroyuki Kuriyama. 2014. Your activity tracker knows when you quit smoking. In *Proceedings of the 2014 ACM international symposium on wearable computers*. 107–110.

[35] Hyoung-Gook Kim, Gee Yeun Kim, and Jin Young Kim. 2019. Music recommendation system using human activity recognition from accelerometer data. *IEEE Transactions on Consumer Electronics* 65, 3 (2019), 349–358.

[36] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[37] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[38] Amanda E Krause and Adrian C North. 2018. 'Tis the season: Music-playlist preferences for the seasons. *Psychology of Aesthetics, Creativity, and the Arts* 12, 1 (2018), 89.

[39] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*. 208–211.

[40] Rita Laureanti, Marco Bilucaglia, Margherita Zito, Riccardo Circi, Alessandro Fici, Fiamma Rivetti, Riccardo Valesi, C Oldrini, Luca T Mainardi, and Vincenzo Russo. 2020. Emotion assessment using Machine Learning and low-cost wearable devices. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 576–579.

[41] Jiayu Li, Weizhi Ma, Min Zhang, Pengyu Wang, Yiqun Liu, and Shaoping Ma. 2021. Know Yourself: Physical and Psychological Self-Awareness With Lifelog. *Frontiers in Digital Health* (2021), 96.

[42] Jiayu Li, Hantian Zhang, Zhiyu He, Rongwu Xu, Pingfei Wu, Min Zhang, Yiqun Liu, and Shaoping Ma. 2022. LifeRec: A Mobile App for Lifelog Recording and Ubiquitous Recommendation. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. 342–346.

[43] Yuzhong Lin, Joran Jessurun, Bauke De Vries, and Harry Timmermans. 2011. Motivate: Towards context-aware recommendation mobile system for healthy living. In *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. IEEE, 250–253.

[44] Hao Liu, Jun Hu, and Matthias Rauterberg. 2009. Music playlist recommendation based on user heartbeat and music preference. In *2009 International Conference on Computer Technology and Development*, Vol. 1. IEEE, 545–549.

[45] Hongyu Lu, Weizhi Ma, Min Zhang, Maarten de Rijke, Yiqun Liu, and Shaoping Ma. 2021. Standing in Your Shoes: External Assessments for Personalized Recommender Systems. (2021).

[46] Yifei Ma, Balakrishnan Narayanaswamy, Haibin Lin, and Hao Ding. 2020. Temporal-Contextual Recommendation in Real-Time. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2291–2299.

[47] David W McDonald. 2003. Ubiquitous recommendation systems. *Computer* 36, 10 (2003), 111–112.

[48] Cédric S Mesnage, Asma Rafiq, Simon Dixon, and Romain P Brixtel. 2011. Music discovery with social networks. In *Workshop on Music Recommendation and Discovery*. Association for Computing Machinery New York, 1–6.

[49] Christos Mettouris and George A Papadopoulos. 2014. Ubiquitous recommender systems. *Computing* 96, 3 (2014), 223–257.

[50] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. 2019. Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–21.

[51] Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperli. 2020. An emotional recommender system for music. *IEEE Intelligent Systems* 36, 5 (2020), 57–68.

[52] YV Srinivasa Murthy and Shashidhar G Koolagudi. 2018. Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review. *ACM Computing Surveys (CSUR)* 51, 3 (2018), 1–46.

[53] Shahriar Nirjon, Robert F Dickerson, Qiang Li, Philip Asare, John A Stankovic, Dezhi Hong, Ben Zhang, Xiaofan Jiang, Guobin Shen, and Feng Zhao. 2012. Musicalheart: A hearty way of listening to music. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. 43–56.

[54] Sungkyu Park, Marios Constantinides, Luca Maria Aiello, Daniele Quercia, and Paul Van Gent. 2020. Wellbeat: A framework for tracking daily well-being using smartwatches. *IEEE Internet Computing* 24, 5 (2020), 10–17.

[55] Florian Resatsch, Stephan Karpischek, Uwe Sandner, and Stephan Hamacher. 2007. Mobile sales assistant: NFC for retailers. In *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services*. 313–316.

[56] Noveen Sachdeva, Kartik Gupta, and Vikram Pudi. 2018. Attentive neural architecture incorporating song features for music recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 417–421.

[57] Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. 2018. Employing consumer wearables to detect office workers' cognitive load for interruption management. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–20.

[58] Markus Schedl, Christine Bauer, Wolfgang Reisinger, Dominik Kowald, and Elisabeth Lex. 2020. Listener Modeling and Context-Aware Music Recommendation Based on Country Archetypes. *Frontiers in Artificial Intelligence* 3 (2020).

[59] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2022. Music recommendation systems: Techniques, use cases, and challenges. In *Recommender Systems Handbook*. Springer, 927–971.

[60] Tiancheng Shen, Jia Jia, Yan Li, Yihui Ma, Yaohua Bu, Hanjie Wang, Bo Chen, Tat-Seng Chua, and Wendy Hall. 2020. Peia: Personality and emotion integrated attentive model for music recommendation on social media platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 206–213.

[61] Yading Song. 2016. *The Role of Emotion and Context in Musical Preference.* Ph. D. Dissertation. Queen Mary University of London.

[62] Zhu Sun, Jie Yang, Jie Zhang, Alessandro Bozzon, Long-Kai Huang, and Chi Xu. 2018. Recurrent knowledge graph embedding for effective recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 297–305.

[63] Yuichiro Takeuchi and Masanori Sugimoto. 2009. A user-adaptive city guide system with an unobtrusive navigation interface. *Personal and Ubiquitous Computing* 13, 2 (2009), 119–132.

[64] Robert E Thayer. 1990. *The biopsychology of mood and arousal.* Oxford University Press.

[65] Md Zia Uddin, Mohammed Mehedi Hassan, Ahmed Alsanad, and Claudio Savaglio. 2020. A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare. *Information Fusion* 55 (2020), 105–115.

[66] Aäron Van Den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. In *Neural Information Processing Systems Conference (NIPS 2013)*, Vol. 26. Neural Information Processing Systems Foundation (NIPS).

[67] Felix Von Reischach, Dominique Guinard, Florian Michahelles, and Elgar Fleisch. 2009. A mobile product recommendation system interacting with tagged products. In *2009 IEEE international conference on pervasive computing and communications*. IEEE, 1–6.

[68] Chenyang Wang, Min Zhang, Weizhi Ma, Yiqun Liu, and Shaoping Ma. 2020. Make it a chorus: knowledge-and time-aware item modeling for sequential recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 109–118.

[69] Shanfeng Wang, Maoguo Gong, Yue Wu, and Mingyang Zhang. 2020. Multi-objective optimization for location-based and preferences-aware recommendation. *Information Sciences* 513 (2020), 614–626.

[70] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 950–958.

[71] Xinxi Wang, David Rosenblum, and Ye Wang. 2012. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia*. 99–108.

[72] Felix Weninger, Florian Eyben, Björn W. Schuller, Marcello Mortillaro, and Klaus R. Scherer. [n. d.]. On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. 0 ([n. d.]). https://doi.org/10.3389/fpsyg.2013.00292 Publisher: Frontiers.

[73] Xuhai Xu, Prerna Chikersal, Janine M Dutcher, Yasaman S Sefidgar, Woosuk Seo, Michael J Tumminia, Daniella K Villalba, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. 2021. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.

[74] Yuyu Yin, Lu Chen, Jian Wan, et al. 2018. Location-aware service recommendation with enhanced probabilistic matrix factorization. *IEEE Access* 6 (2018), 62815–62825.

[75] Sunkyung Yoon, Edelyn Verona, Robert Schlauch, Sandra Schneider, and Jonathan Rottenberg. 2020. Why do depressed people prefer sad music? *Emotion* 20, 4 (2020), 613.

[76] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. 2006. Hybrid Collaborative and Content-based Music Recommendation Using Probabilistic Model with Latent User Preferences.. In *ISMIR*, Vol. 6. 296–301.

[77] Eva Zangerle, Martin Pichl, and Markus Schedl. 2018. Culture-aware music recommendation. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. 357–358.

[78] Weifeng Zhang, Jingwen Mao, Yi Cao, and Congfu Xu. 2020. Multiplex Graph Neural Networks for Multi-behavior Recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2313–2316.

[79] Yiwen Zhang, Chunhui Yin, Qilin Wu, Qiang He, and Haibin Zhu. 2019. Location-aware deep collaborative filtering for service recommendation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2019).

[80] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1059–1068.

[81] Zack Zhu, Hector F Satizabal, Ulf Blanke, Andres Perez-Uribe, and Gerhard Tröster. 2015. Naturalistic recognition of activities and mood using wearable electronics. *IEEE Transactions on Affective Computing* 7, 3 (2015), 272–285.

## A QUESTIONNAIRE FOR MUSIC PREFERENCE AND CONTEXT INFORMATION

1. Gender: A. Male B. Female C. Other

2. Age: A. Under 18 B. 18-25 C. 26-30 D. 31-40 E. 41-50 F. 51-60 G. Over 60

3. Highest education: A. Under high school B. High school C. Bachelor D. Master E. Ph.D

4. Occupation: ________________________

Please answer the following questions according to your daily listening habits **in the past month**.

5. The average time you spend listening to music per day is about:

A. never listen to music B. less than 1 hour C. 1-2 hours D. 2-3 hours E. over 3 hours

6. When listening to music, how often do you use the recommendation of the music App (such as daily recommendation, recommended playlist, similar music, etc.)?

A. Never (use less than 2 times in 10 times of listening to music).
B. Occasionally (use 2-3 times in 10 times of listening to music).
C. Sometimes (use 4-6 times in 10 times of listening to music).
D. Often (use 7-9 times in 10 times of listening to music).
E. Always.

7. Please select the music genres you often listen to. (Choose at least 1 item.)

A. Pop B. Classical C. Folk/country
D. Rock/punk/metal E. Hip-hop/rap F. Electronic
G. Light H. Jazz I. Dance J. Others

8. In the process of listening to music, what factors affect your preference for music? Please sort the following factors. (Mark the number in the box.) [11]

□ Current mood.
□ Context information such as weather, brightness, noise, etc.
□ Singer/player.
□ Current activity
□ Music genre.
□ Music content and melody.
□ Music popularity.

In the following questions, please try to imagine yourself in different mood status.

9. How much do you want to start listening to music when you are in the following moods? (rating from 0 to 10. 0 for not at all and 10 for very much)

Frustration/depression: ______
Sadness: ______
Anxiety/anger: ______
Calm: ______
Pleasure: ______
Excitement: ______

10. Would you choose different music in different moods?

A. Always.
B. Often.
C. Sometimes.
D. Occasionally.
D. Never.

11. What music genres would you prefer in the following mood? (Tick under the selected music genres, and choose all music genres you favor.)

| Moods | Pop | Classical | Folk | Rock | Hip-hop | Electronic | Light | Dance | Others | Don’t listen |
|---|---|---|---|---|---|---|---|---|---|---|
| Felt depressed | | | | | | | | | | |
| Felt sad | | | | | | | | | | |
| Felt angry | | | | | | | | | | |
| Felt calm | | | | | | | | | | |
| Felt happy | | | | | | | | | | |
| Felt excited | | | | | | | | | | |

12. What music emotion do you generally choose in the following moods?

| Moods | Sad | Healing | Relaxing | Exciting | Cheerful | Quiet | Others | Don’t listen |
|---|---|---|---|---|---|---|---|---|
| Felt depressed | | | | | | | | |
| Felt sad | | | | | | | | |
| Felt angry | | | | | | | | |
| Felt calm | | | | | | | | |
| Felt happy | | | | | | | | |
| Felt excited | | | | | | | | |

[11] The choices are randomly displayed to the participants in practice.

13. Please choose 'Often' for this question.
A. Always.
B. Often.
C. Sometimes.
D. Occasionally.
D. Never.

14. Do you have different preferences of the same music in different moods?
A. Always.
B. Often.
C. Sometimes.
D. Occasionally.
D. Never.

15. Do you think your moods will change before and after listening to music?
A. Always.
B. Often.
C. Sometimes.
D. Occasionally.
D. Never.

16. What factors affect your mood before and after listening to music? Please sort. (Select at least 4 items and mark the number in the box.)

☐ Music genre.
☐ Music content and melody.
☐ Context information such as weather, brightness, noise, etc.
☐ Previous mood.
☐ Music popularity.
☐ Current activity.
☐ Singer/player.

The following questions refer to your daily listening habits **in the past month**.

17. How often do you listen to music when doing the following activities?

| Activities | Never | Occasionally | Sometimes | Often | Always |
|---|---|---|---|---|---|
| Study/work | | | | | |
| Housework | | | | | |
| Commute | | | | | |
| Entertain | | | | | |
| Exercise | | | | | |
| Rest | | | | | |
| Eat | | | | | |
| Sleep | | | | | |

18. How important is music to you when you are doing the following activities? (1:Totally unimportant. 4: Neutral. 7: Very important.)

| Activities | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Study/work | | | | | | | |
| Housework | | | | | | | |
| Commute | | | | | | | |
| Entertain | | | | | | | |
| Exercise | | | | | | | |
| Rest | | | | | | | |
| Eat | | | | | | | |
| Sleep | | | | | | | |

19. What is the main purpose of listening to music when you are doing the following activities?

A. Reduce boredom: use music to reduce boredom and allocate extra attention.

B. Focus: maintain attention and alertness to current activities.

C. Regulate emotion: regulate the current emotional state.

D. Guide activities: the movement of the activity is synchronized with the rhythm of the music, which makes the activity more rhythmic/easier to proceed.

E. Aesthetic enjoyment: listen to music in order to appreciate and enjoy it.

| Activities | A | B | C | D | E | Don't listen |
|---|---|---|---|---|---|---|
| Study/work | | | | | | |
| Housework | | | | | | |
| Commute | | | | | | |
| Entertain | | | | | | |
| Exercise | | | | | | |
| Rest | | | | | | |
| Eat | | | | | | |
| Sleep | | | | | | |

20. What music genres do you prefer when doing the following activities? (Please select all you favor)

| Moods | Pop | Classical | Folk | Rock | Hip-hop | Electronic | Light | Dance | Others | Don't listen |
|---|---|---|---|---|---|---|---|---|---|---|
| Activities | | | | | | | | | | |
| Study/work | | | | | | | | | | |
| Housework | | | | | | | | | | |
| Commute | | | | | | | | | | |
| Entertain | | | | | | | | | | |
| Exercise | | | | | | | | | | |
| Rest | | | | | | | | | | |
| Eat | | | | | | | | | | |
| Sleep | | | | | | | | | | |

## B PARTICIPANT INHERENT MUSIC PREFERENCE ANALYSIS

In Section 5, a series of analyses are conducted on all participants in general. Nevertheless, different participants may have distinct influences of context and mood on music preference due to individual heterogeneity. Therefore, we explore the effect of individual inherent music preference on context-related music ratings in this section.

Participants' inherent music preferences may affect context-related music ratings. For instance, comparing preferences for folk music when participants were in different activities, we found music ratings significantly

Table 5. The statistical results of four GLMs with different music genres and context categories. $\beta_{c,i}$, $\beta_p$, and $\beta_{m,i}$ are coefficients of context, preference, and their interactions. (r) indicates the reference type of the category variable, i.e., context type. **Bold** fonts represent significant coefficients (p<0.05).

| **Genre** | **Context cat.** | **Context type** | $\boldsymbol{\beta_{c,i}}$ | **p value** | $\boldsymbol{\beta_{m,i}}$ | **p value** | $\boldsymbol{\beta_p}$ | **p value** |
|---|---|---|---|---|---|---|---|---|
| rock | activity | sleeping (r) | 0 | - | 0 | - | 0.248 | 0.118 |
| | | standing | 0.154 | 0.652 | -0.100 | 0.226 | | |
| | | sitting | -0.035 | 0.877 | -0.071 | 0.728 | | |
| | | walking | **-0.924** | 0.000 | 0.329 | 0.200 | | |
| folk | activity | sleeping (r) | 0 | - | 0 | - | 0.123 | 0.421 |
| | | standing | 0.667 | 0.314 | 0.083 | 0.827 | | |
| | | sitting | -0.233 | 0.698 | 0.550 | 0.108 | | |
| | | walking | **-1.374** | 0.032 | 0.203 | 0.583 | | |
| folk | time | 8 a.m. - 12 p.m. (r) | 0 | - | 0 | - | 0.133 | 0.632 |
| | | 12 p.m.-18 p.m. | -0.735 | 0.410 | 0.240 | 0.626 | | |
| | | 18 p.m. - 1 a.m. | **-0.644** | 0.034 | 0.057 | 0.885 | | |
| pop | weather | sunnry (r) | 0 | - | 0 | - | 0.395 | 0.051 |
| | | cloudy | 0.008 | 0.992 | 0.196 | 0.656 | | |
| | | rainy | 0.163 | 0.836 | 0.105 | 0.813 | | |

lower when walking (as shown in Figure 8). But this may not necessarily be the case for some participant who likes or dislikes folk music. Hence, we analyze the mixed effects of context and inherent preference on music ratings with the records in the field study, as well as participants' inherent preference for each genre (*dislike*, *neutral*, or *like*, collected at the end of the field study).

In the previous analysis, Figure 8 shows average ratings on typical genres in different contexts, where exist four significant situations: folk and rock music in different activities, folk music in different time periods, and pop music in different weather. We analyze with a generalized linear model (GLM) for each of the above four situations. Taking the ratings of folk music in different activities as an example, we fit the following equation for all the field study records with folk music:

$$R = \sum_{i=1,2,\ldots,C_n-1} (\beta_{c,i} + \beta_{m,i}P) \cdot C_i + \beta_p P + \beta_0 + e \tag{8}$$

Where $P$ is a continuous variable for self-report preference on folk music (0 for dislike, 1 for neutral, and 2 for like). $C$ is a category variable indicating types of context: sleeping, standing, sitting, or walking. $C_n$ is the number of context types. And $R$ is the 5-score rating. $\beta_{c,i}$, $\beta_p$, $\beta_{m,i}$ are coefficients corresponding to effects of context, preference, and their interactions. $\beta_0$ is the general intercept, and $e$ is the general residual error. Note that the category variable in GLM has only $C_n - 1$ degrees of freedom, so the numbers of $\beta_{c,i}$ and $\beta_{m,i}$ are both $C_n - 1$. Similar GLMs are built for the other three situations.

We report the statistical results of the coefficients in four GLMs in Table 5, which shows the following findings: Firstly, $\beta_{c,i}$ indicates that when considering inherent preference, context features still help the prediction in the same way as in Figure 8. And the promotion is significant for rock and folk while walking and folk from 6 p.m. to 1 a.m. Hence, regardless of participants' inherent preference, context features may affect music preference. Secondly, inherent preference has a positive impact on the result with $\beta_p > 0$, which means that the rating is high when the participant prefers the type of music inherently and vice versa. It indicates that self-report preference is consistent with data collected in the field study. Furthermore, coefficient $\beta_{m,i}$ shows the effect

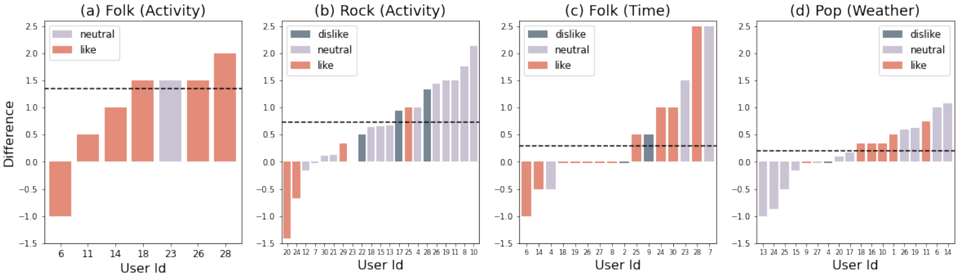


Fig. 17. Differences of music rating between context types in the same category and music genre for each participant. The black dashed line indicates average difference of all participants.

of inherent self-report preference in each context. In the different contexts of the same category, the effect of inherent self-report preference is different, but they are all not significant. Due to the limited size of the data set, the amount of data divided into every context type and the music genre is small, and many results are not significant. But the significance of $\beta_{c,i}$ indicates that whatever inherent preference is, context features influence preference in many cases.

To present an individual-level comparison more intuitively, we analyze the *difference* in ratings of different context types in four genre contexts for each participant. *Difference* is defined according to the context category. For the activity category, because Figure 8 reveals that rating is lower when walking than in the other three activities, the definition of *difference* is

$$(R(sleeping) + R(standing) + R(sitting))/3 - R(walking) \tag{9}$$

Where $R(\cdot)$ is the mean value of ratings in category $\cdot$ under the music genre. For the other two conditions, because the preference of different context types is different for folk of time category and pop of weather category, the definition of difference for each category is

$$R_{max} - R_{min} \tag{10}$$

Where $R_{max}$ is the context with maximal rating, i.e., 8 a.m.-12 p.m. and cloudy day, respectively. $R_{min}$ is the context with minimal rating, i.e., 6 p.m.-1 a.m. and sunny day, respectively. Generally, a higher difference indicates that participant's preference is more sensitive to context changes than average, and a lower means the opposite.

We report the statistical results of differences in Figure 17. For Figure 17(a) and Figure 17(b), we retain participants who have listened to the music of this genre in a *walking* state and at least one of the remaining three context types. For the last two conditions, *Folk (Time)* and *Pop (Weather)*, we retain participants who have records in context with both maximal and minimal ratings.

Figure 17(d) indicates that the difference between subjects who like the pop genre is higher than average (black dashed). Whereas, in the remaining three figures, the difference between participants who like/dislike folk or rock genre has a wide distribution. Hence, participants who prefer the pop genre give a much higher rating on a cloudy day than on a sunny day, i.e., they are more sensitive to weather, and participants who dislike the pop genre are less sensitive to weather. However, there are no consistent changes in other music genres and contexts. Especially, users who dislike some music genres show preference differences between contexts. For example, User 18, User 17, and User 28 all dislike rock music (in Figure 17(b)), but their differences between activities are all greater than zero (for User 17 and User 28, greater than average). In Figure 17(c), although not prefer folk, User 9 also shows an above-average difference in preferences changing with time.

In conclusion, there is no consistent and significant difference in the effect of inherent preference on the context. Therefore, we do not use the self-report inherent preferences as part of the inputs for our method, because they are not explicitly associated with the subject's context-related music preference change. Besides, self-reports are usually not available in recommendation systems. Instead, following the practices of traditional recommender systems, we utilize historical preference to predict musical ratings and achieve encouraging results in experiments in Section 6.

Nevertheless, the influence of heterogeneous individual-level characteristics is an interesting but complex topic, and our discussions is only a preliminary exploration. We will continue to explore the problem of individual heterogeneity in the future.